­­­­modelos de regressão logistica multinomial

A estatística é utilizada nas mais diversas áreas do conhecimento científico com a pretensão de responder a problemas subjacentes a estas. Em muitos estudos estatísticos pretende-se saber o efeito que determinadas variáveis provocam na variável resposta (interesse), acreditando que a variabilidade desta é explicada pelas outras.

Modelo de Regressão Multinominal

Através da Regressão Multinominal pode identificar-se variáveis explicativas que são significativas e procurar o modelo final que seja o mais parcimonioso, não perdendo valor explicativo. Para reduzir o efeito das categorias de algumas variáveis que possuíam frequência muito baixa (penalizando a estimação do modelo), tornou-se necessário reagrupar a variável Idade em novas categorias: grupo 1 [65, 75]; grupo 2 [76, 85] e grupo 3 [86, 95]. Pelo mesmo motivo a variável escala de depressão foi também reagrupada em 6 novas categorias.

o modelo de regressão logística multinominal, devido à natureza qualitativa (nominal) da variável com três categorias. Este tipo de regressão é reconhecidamente mais apropriado por apenas considerar a relação entre as variáveis “independente” e “dependente” excluindo a relação entre as variáveis independentes (Cramer, 2003; Maroco, 2007).

Livro:

Em resumo, o ajuste e a interpretação dos resultados de um modelo de regressão logística multinominal seguem o mesmo paradigma básico que foi seguido para um modelo binário.

A diferença é que o utilizador deve estar ciente da possibilidade de que declarações comparativas informativas possam ser necessárias para as razões de chances múltiplas para cada covariável.

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias. No presente estudo foi empregado um tipo de regressão logística denominada multinomial, a qual permite a análise das variáveis explicativas com a variável resposta em seus vários níveis, cuja variável dependente possui mais de duas classes mutuamente exclusivas11.

(<https://www.researchgate.net/publication/305038774_Modelo_de_decisao_sobre_o_uso_de_preservativos_Uma_regressao_logistica_multinomial>

)

<https://www.researchgate.net/publication/280742156_Analise_probabilistica_da_condicao_de_degradacao_dos_revestimetos_petreos_fixados_directamente_ao_suporte>

Tradicionalmente, esses resultados de pesquisa eram tratados por regressão de mínimos quadrados ordinários (OLS) ou análise de função discriminante linear (Hosmer & Lemeshow, 2000). Entretanto, como resultado de sua natureza, ambas as técnicas dependem de premissas estatísticas rigorosas, portanto normalidade de variáveis ​​independentes, linearidade de relacionamentos, multicolinearidade entre variáveis ​​independentes, matrizes de dispersão iguais para análise discriminante (Tabachnick et al., 2001). Essas suposições que não são facilmente observadas em um ambiente dinâmico fazem parte da regressão múltipla 2.

A introdução da regressão logística multinomial foi uma análise de regressão alternativa para atender às condições que não necessariamente obedecem às premissas listadas acima, com exceção da multicolinearidade (Hosmer & Lemeshow, 2000). Na última década, a técnica, como outros métodos de análise de dados univariados e multivariados, começou a encontrar um lugar de destaque nas indústrias de medicina, engenharia e manufatura. Esse desenvolvimento levou os pesquisadores em análise de risco a construir modelos estatísticos mais precisos e úteis, aplicando-os na análise de risco (Liebenberg e Hoyt, 2003; Hosmer e Lemeshow, 2000). Enquanto isso, apesar de estar em uso na análise estatística geral por muitos anos, recebeu pouca atenção na literatura de análise de risco em comparação com outras regressões

aplicações relacionadas à modelagem de variáveis ​​explicativas e de resposta (Liebenberg e Hoyt, 2003; Hosmer e Lemeshow, 2000).

O modelo de regressão logística multinomial (politômica) é uma extensão simples do modelo de regressão logística binomial. É usado quando a variável dependente possui mais de duas categorias nominais ou não ordenadas, nas quais a codificação fictícia (3) de variáveis ​​independentes é bastante comum. Ao usar a regressão logística multinomial na análise de risco, a variável dependente (resposta) é dummy codificada em várias variáveis ​​1/0 (consulte a seção 3 para obter detalhes).

Isso significa que existe uma variável para todas as categorias, exceto uma, portanto, se houver categorias M, haverá variáveis ​​dummy M-1. Apenas uma categoria possui sua própria variável dummy. A variável dummy de cada categoria tem um valor de 1 para sua categoria e um 0 para todas as outras. Uma categoria, a categoria de referência, não precisa de sua própria variável dummy, uma vez que é identificada exclusivamente por todas as outras variáveis ​​sendo 0. Com relação ao exposto acima, o analista de risco que usa regressão logística multinomial pode estimar um modelo de regressão logística binária separado para cada uma dessas variáveis ​​fictícias. O resultado são modelos de regressão logística binária M-1. O fator mais significativo a ser considerado aqui é que cada um informa o efeito dos preditores de risco na probabilidade de sucesso nessa categoria, em comparação com a categoria de referência. Observando, porém, que cada modelo possui seus próprios coeficientes de interceptação e regressão - a razão é que os preditores dos processos de análise de risco podem afetar cada categoria de maneira diferente (consulte as seções 2 e 3 para obter detalhes).

(<https://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=0&sid=6a98f215-e2df-4813-8162-e9eb3bb12605%40sessionmgr4006>

)

McFadden em 1974 propôs uma mudança do modelo de regressão logística e chamou-lhe discret choice model. Como resultado o modelo é normalmente referido como discrete choice model na literatura de negócios e economia enquanto que é chamado de multinomial, poliatômico na saúde e ciências da vida. Pode-se usar uma variável resposta com qualquer número de níveis para ilustrar a extensão do modelo e métodos. No entanto, os detalhes são mais facilmente ilustrados com três categorias. Uma generalização adicional para mais de três categorias é mais um problema de notação do que conceção.

A regressão logística multinomial trata de um modelo de regressão logística que permite que a variável categórica dependente apresente mais de duas categorias, permitindo que sejam de natureza ordinal ou nominal. Uma das categorias da variável dependente deve ser escolhida como referência.

ou ainda, mas menos utilizado, regressão logística policotómica.

Tendo em consideração todos os aspetos apresentados

Através do uso da estatística, introduzir um alarme nas linhas de produção, sempre que a probabilidade de falha ultrapasse determinado valor.

o impacto da aplicação de algoritmos que permitam resolver problemas de escalonamento da produção em contexto industrial, através da otimização de algumas medidas de desempenho.

Fast Estimation of Multinomial Logit Models: R Package mnlogit ((Hasan, Zhiyu, & Mahani, 2016))

Journal of Statiscal Software

O pacote de R mnlogit pode ser usado para estimar modelos de regressão logística multinomial, especialmente os que envolvem um grande número de categorias e variáveis.

Comparando com o software existente o mnlogit poderá ser 10 a 50 vezes mais rápido em problemas de tamanho considerável e mais de 100 vezes para problemas maiores. O mnlogit alcança sua eficiência computacional acelerando drasticamente o cálculo da matriz Hessian da função log-verossimilhança através da exploração da estrutura em matrizes que surgem em cálculos intermediários. Essa exploração eficiente de estruturas de dados intermediárias permite que o mnlogit utilize a memória do sistema com muito mais eficiência, de modo que, para a maioria das aplicações, o mnlogit exija menos memória que o software comparável por um fator proporcional ao número de categorias de modelo.

O Modelo de regressão logística multinomial é uma extensão simples do modelo de regressão logística binária. Este modelo foi proposto por McFadden em 1974 e consistiu numa alteração do modelo de regressão logística ao qual lhe chamou discret choice model. Esta nomenclatura é utilizada no mundo dos negócios enquanto que na vida das ciências é utilizado o termo multinomial. Este será o termo usado nesta dissertação. Também é utilizado em machine learning como técnica de classificação linear (Hastie, Tibshirani, & Friedman, 2009)

Treinar modelos é um desafio computacional para efetuar o calculo de um grande número de coeficientes, que aumenta linearmente com o número de categorias e o número de variáveis. Apesar dos modelos logit multinomiais serem caros para efetuar estimativas

Segundo (Hasan et al., 2016) o pacote *mnlogit* fornece a capacidade de lidar com colinearidade de dados, além de ter a vantagem de efetuar cálculos muito rápidos. O mnlogit, utiliza a estimativa de máxima verossimilhança (MLE) usando o método de Newton-Raphson (NR) em que foi explorada a estrutura e dispersão de matrizes de dados intermédias para obter cálculos mais rápidos da função Hessian de log-verossimilhança (Train, 2003).

Bibliografia

Hasan, A., Zhiyu, W., & Mahani, A. S. (2016). Fast estimation of multinomial logit models: R package mnlogit. *Journal of Statistical Software*, *75*(3). https://doi.org/10.18637/jss.v075.i03

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statiscal Learning : Data Mining, Inference, and Prediction* (Second). Springer.

Train, K. E. (2003). Discrete choice methods with simulation. *Discrete Choice Methods with Simulation*, *9780521816960*, 1–334. https://doi.org/10.1017/CBO9780511753930