CE073 - Análise de Dados Categóricos
Departamento de Estatística
Universidade Federal do Paraná

Trabalho No.2

11 de setembro de 2023

Redigir de maneira individual e entregar na área correspondente no sistema **Microsoft Teans** um relatório eletrônico até o dia **02 de outubro de 2023**.

Capítulo 2- Regressão com resposta binária

1- A falha de um O-ring nos foguetes de reforço do ônibus espacial Challenger levou à sua destruição em 1986. Usando dados de lançamentos anteriores de ônibus espaciais, Dalal et al. (1989) examinaram a probabilidade de uma falha do O-ring em função da temperatura no lançamento e da pressão de combustão. Os dados de seu artigo estão incluídos no arquivo:

challenger = read.csv(file = "http://leg.ufpr.br/~lucambio/CE073/20222S/challenger.csv")

Descrição das variáveis:

- Flight: número do voo
- Temp: Temperatura (em graus Fahrenheit) no lançamento
- Pressure: Pressão de combustão (psi)
- O.ring: Número de falhas de O-ring do campo primário
- Number: Número total de O-rings de campo primário (seis no total, três cada para os dois foguetes de reforço)

A variável de resposta é O.ring e as variáveis explicativas são Temp e Pressure.

Responda o seguinte:

a) Os autores usam a regressão logística para estimar a probabilidade de um O-ring falhar. Para usar este modelo, os autores precisaram assumir que cada O-ring é independente para cada lançamento. Discuta por que essa suposição é necessária e os possíveis problemas com ela. Observe que uma análise posterior ajudou a aliviar as preocupações dos autores sobre a independência.
b) Estime o modelo de regressão logística usando as variáveis explicativas de forma linear.
c) Realize testes da razão de verossimilhanças (LRTs) para julgar a importância das variáveis explicativas no modelo.

d) Os autores optaram por remover Pressure do modelo com base nos LRTs. Com base em seus resultados, discuta por que você acha que isso foi feito. Há algum problema potencial com a remoção dessa variável?

2- Continuando o Exercício 1, considere o modelo simplificado

$$logit(\pi) = \beta_0 + \beta_1 \, \text{Temp},$$

onde π é a probabilidade de falha do O-ring. Responda o seguinte:

a) Estime o modelo

b) Construa dois gráficos: (1) π vs. Temp e (2) Número esperado de falhas vs. Temp. Use uma faixa de temperatura de 31°F a 81°F no eixo x, mesmo que a temperatura mínima no conjunto de dados seja 53°F.

c) Inclua as bandas do intervalo de confiança de Wald de 95% para no gráfico. Por que as bandas são muito mais largas para temperaturas mais baixas do que para temperaturas mais altas?

d) A temperatura era 31°F no lançamento do Challenger em 1986. Estime a probabilidade de uma falha do O-ring usando esta temperatura e calcule um intervalo de confiança correspondente. Discuta quais suposições precisam ser feitas para aplicar os procedimentos de inferência.

d) Em vez de usar intervalos Wald ou LR perfilada para a probabilidade de falha, Dalal et al. (1989) usam um bootstrap paramétrico para calcular intervalos. Seu processo foi:

(1) simular um grande número de conjuntos de dados, por exemplo 500 ou 1000 com $n = 23$ para cada, a partir do modelo estimado de $logit(\widehat{\pi}) = \widehat{\beta}_0 + \widehat{\beta}_1 \, \text{Temp}$;

(2) estimar novos modelos para cada conjunto de dados, digamos $logit(\widehat{\pi}^*) = \widehat{\beta}_0^* + \widehat{\beta}_1^* \, \text{Temp}$ e

(3) calcular $\widehat{\pi}^*$ a uma temperatura específica de interesse. Os autores usaram os quantis observados de 0.05 e 0.95 da distribuição simulada de $\widehat{\pi}^*$ como seus limites de intervalo de confiança de 90%.

Usando o bootstrap paramétrico, calcule intervalos de confiança de 90% separadamente em temperaturas de 31°F e 72°F. Os métodos bootstrap usados aqui correspondem ao que é conhecido como intervalo percentil. Existem maneiras melhores de formar intervalos de confiança usando o bootstrap, ou seja, o BCa e os intervalos estudantis. Por favor, veja Davison and Hinkley (1997) para uma discussão e as funções boot() e boot.ci() no pacote boot para implementação.

f) Determine se um termo quadrático é necessário no modelo para a temperatura.

3- Continuando o Exercício 2:

a) Investigue se a probabilidade estimada de falha teria mudado significativamente se um modelo de regressão probit ou log-log complementar fosse usado em vez de um modelo de regressão logística.

b) Investigue também se aplicar uma função de ligação paramétrica aprimora a probabilidade de falha.

4- Turner et al. (1992) usam a regressão logística para estimar a taxa na qual o picloram, um herbicida, mata o larkspur, uma erva daninha. Seus dados foram coletados pela aplicação de quatro níveis diferentes de picloram em parcelas separadas e o número de ervas daninhas mortas dentro da parcela foi registrado.

Os dados estão no arquivo

dados = read.csv(file = "http://leg.ufpr.br/∼lucambio/ADC/picloram.csv")

O experimento foi realizado em três execuções separadas, que são indicadas pela variável rep no arquivo de dados. Os autores estimam um modelo de regressão logística sem levar em conta a variável rep e nós o mesmo aqui.

a) Estime um modelo de regressão logística usando a quantidade de picloram como variável explicativa e o número de ervas daninhas mortas como variável de resposta.

b) Trace a proporção observada de ervas daninhas mortas e o modelo estimado. Descreva o quão bem o modelo se ajusta aos dados. Investigue também se aplicar uma função de ligação paramétrica aprimora a probabilidade de falha.

c) Estime o nível de taxa de morte de 0.9 (?LD90?) para o picloram. Adicione linhas ao gráfico em (b) para ilustrar como ele é encontrado, a função **segment()** pode ser útil para esse propósito. Na Seção http://leg.ufpr.br/ lucambio/ADC/ADC02.html#2.4, o exercício 21 fornece uma explicação da definição estimação da dose letal x_π, comumente chamado de ?LDz?.