O método Jackknife

Fundamentos e aplicações

Prof. Walmes Zeviani
walmes@ufpr.br

Laboratório de Estatística e Geoinformação
Departamento de Estatística
Universidade Federal do Paraná

Atualizado em 2018-09-03

Justificativas

- Métodos computacionalmente intensivos para inferência estatística são usados quando as abordagens tradicionais não são adequadas.
  - Resultados assintóticos em pequenas amostras.
  - Violação de pressupostos.
  - Não existência de mecanísmos de inferência específicos.
- Tais métodos se baseiam em reamostragem e/ou simulação.
- Podem ser aplicados em muitos contextos.

Objetivos

- Apresentar a ideia principal sobre o método Jackknife.
- Ilustrar propriedades com aplicações.

- O método Jackknife foi proposto por Quenouille 1956.
- Jackknife refere-se a um canivete suiço, fácil de carregar e de várias utilidades.
- Devido a isso, Tukey 1958 cunhou o termo em Estatística como uma abordagem geral para testar hipóteses e calcular intervalos de confiança.

Amostras Jackknife

As amostras Jackknife são definidas por deixar k observações de fora do conjunto observado $x = \{x_1, x_2, \ldots, x_n\}$ por vez.

O caso mais prático e comum é quando $k = 1$. Assim a amostra Jackknife i é definida como

$$x_{(i)} = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}, \tag{1}$$

com $i = 1, 2, \ldots, n$.

- O tamanho de cada amostra Jackknife é $m = n - k$.
- O número de amostras distintas é $\binom{n}{k}$.
- No caso de $k = 1$, denota-se por $\{x_{(i)}\}$ com $i = 1, \ldots, n$.
- As amostras são obtidas de forma sistemática, portanto, trata-se de uma abordagem determinística.

A intuição do método

A ideia é fundamentada no estimador da média

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2}$$

A média com a j-ésima observação removida é calculada por

$$\bar{X}_{-j} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i\right) - X_j. \tag{3}$$

Combinando as expressões anteriores, pode-se determinar o valor de X_j por

$$X_j = n\bar{X} - (n-1)\bar{X}_{-j}. \tag{4}$$

Essa expressão não tem valor prático para o caso da média, porém tem utilidade para outras estatísticas.

Definição

Suponha que θ seja um parâmetro a ser estimado a partir de uma função dos dados

$$\hat{\theta} = f(X_1, X_2, \ldots, X_n). \tag{5}$$

A estimativa em cada amostra Jackknife é aquela obtida deixando k observações de fora por vez. No caso de $k = 1$, é definida por

$$\hat{\theta}_{-i} = f(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) = f(X_{(i)}), \tag{6}$$

é chamada de estimativa parcial.

A quantidade

$$\theta_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{-i} \tag{7}$$

é denominada de pseudo-valor e se baseia nas diferenças ponderadas da estimativa com todas as observações ($\hat{\theta}$) e na estimativa parcial, ou seja, aquela sem a i-ésima observação ($\hat{\theta}_{-i}$).

Estimativa pontual e variância

O estimador pontual de Jackknife é definido por

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} \theta_{(i)}, \tag{8}$$

ou seja, é a média dos pseudo-valores.

ATENÇÃO: Os valores $\hat{\theta}$ e $\tilde{\theta}$ não são necessariamente iguais nos casos gerais.

Se for assumido que os valores $\theta_{(i)}$, $i = 1, \ldots, n$, são independentes, a variância do estimador de Jackknife e dada por

$$\text{Var}(\tilde{\theta}) = \frac{S^2_{\theta}}{n}, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{(i)} - \tilde{\theta})^2. \tag{9}$$

Detalhes

De acordo com Efron e Hastie 2016

- É um procedimento não paramétrico pois nenhuma suposição é feita sobre a distribuição dos dados.
- É facilmente automatizável. Um único algoritmo pode ser escrito tendo como argumentos a amostra e a estatística de interesse $f(.)$.
- É determinístico, portanto, toda execução do procedimento irá fornecer os meus resultados.
- Existe a suposição implicita de comportamento suave da função f em relação a cada elemento da amostra Jackknife.
- O erro-padrão de Jackknife é viciado para estimar o verdadeiro erro padrão, pois os pseudo-valores não são independentes.

Próxima aula

- Implementação e aplicações de Jackknife.

Avisos

- Sabatina estará disponível a partir de Qua.

Referências bibliográficas

- Efron, Bradley e Trevor Hastie (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science (Institute of Mathematical Statistics Monographs)*. Cambridge University Press.
- Quenouille, M. H. (1956). "Notes on Bias in Estimation". Em: *Biometrika* 43.3/4, p. 353.
- Tukey, John W. (1958). "Bias and confidence in not quite large samples (abstract)". Em: *The Annals of Mathematical Statistics* 2.29, p. 614.