Estimador de Máxima Verossimilhança
Estudo de Caso - Regressão Poisson

Wagner Hugo Bonat - LEG/DEST, UFPR [1]

Resumo:
Este texto descreve de forma rápida o processo de estimação baseado em Verossimilhança para um modelo de regressão Poisson. **Palavras-chave**: *Poisson, Verossimilhança, Perfilhada..*

1 Introdução

O número de vezes que um determinado evento ocorre é uma forma comum de dados. Exemplos de dados de contagens ou frequências são: número de pessoas infectadas por uma doença, número de chamadas telefônicas que chegam a uma central, número de assaltos em uma cidade, entre outros.

A distribuição de Poisson $P(\mu)$ é muito usada para modelar dados de contagens. Se Y é o número de ocorrências, sua distribuição de probabilidade pode ser escrita como

$$f(y) = \frac{\mu^y \exp^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots$$

onde μ é o número médio de ocorrências. Pode ser mostrado que $E(Y) = \mu$ e $Var(Y) = \mu$.

Considere por exemplo que o número de defeitos em chapas de aço fabricadas por uma determinada fábrica segue uma distribuição de Poisson. Porém, esta fábrica produz chapas com diferentes tamanhos, por diferentes operadores, em diferentes turnos e máquinas. Estas características podem estar relacionadas ao número de defeitos encontrados em uma determinada chapa de aço. Estas características dentro de um modelo de regressão são denominadas de variáveis explanatórias, explicativas ou covariáveis. O objetivo do modelo de regressão Poisson é modelar o relacionamento destas possíveis covariáveis com a média da variável resposta.

No modelo de regressão Poisson o efeito das covariáveis X na resposta Y é modelado através do parâmetro μ. A forma mais comum de tal relacionamento é $g(\mu) = \beta_0 + \beta_1 X$, também chamado de preditor linear. O objetivo deste trabalho é estimar os parâmetros envolvidos em um modelo de regressão Poisson, bem como, ilustrar por um processo de simulação suas propriedades assintóticas. Será feita avaliação para o vício, nível de cobertura dos intervalos de confiança do tipo Wald e serão obtidos intervalos por perfil de Verossimilhança.

[1] Contato: *wagner@leg.ufpr.br.*

2 Regressão Poisson

Seja $Y_1, \ldots, Y_n$ variáveis aleatórias independentes com Y_i denotando o número de eventos observados em n_i expostos. Por exemplo, em uma cidade onde a população é de $100,000$ observa-se 1500 pessoas com certa doença. O número esperado para Y_i pode ser escrito como $E(Y_i) = \mu_i = n_i\theta_i$.

Outro exemplo, suponha que Y_i é o número de indenizações pagas por uma seguradora para uma particular marca e modelo de carro. Isto vai depender do número de carros deste tipo que estão sendo segurados, n_i, e outras covariáveis que afetam θ_i, tal como, a idade do carro e a localização onde esta sendo usado. O subescrito i é usado para denotar diferentes combinações de marcas, modelos, idade, localização e assim por diante.

A dependência de θ_i nas covariáveis é usualmente modelada por

$$\theta_i = \exp^{x_i^T\beta}$$

.

Por isso, o modelo linear generalizado é

$$E(Y_i) = \mu_i = n_i \exp^{x_i^T\beta}; \quad Y_i \sim P(\mu_i)$$

.

A função de ligação natural é a função logaritmo

$$\log \mu_i = \log n_i + x_i^T\beta$$

O termo $\log n_i$ é chamado de **offset**, neste trabalho em particular será assumido que $n_i = 1$ e apenas uma covariável X_1 em um modelo com intercepto, por simplicidade. Sendo assim, a especificação do modelo está completa.

3 Inferência baseado na Verossimilhança - Regressão Poisson

Para ilustrar o modelo em um primeiro momento vamos simular um conjunto de dados proveniente deste modelo.

```
> x <- seq(1, 10, l = 100)
> b0 = 0.5
> b1 = 0.3
> lambda <- exp(b0 + b1 * x)
> set.seed(123)
> y <- rpois(100, lambda = lambda)
```

Um diagrama de dispersão nos auxilia a entender o modelo.

Como em todos os casos de estimação por Máxima Verossimilhança o primeiro passo é encontrar a função de Verossimilhança. Neste caso, tem-se

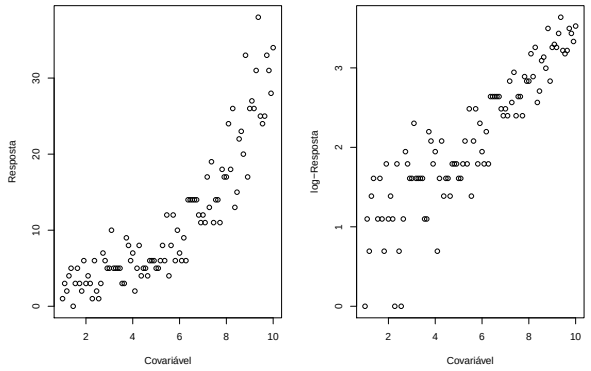

Figura 1: Diagrama de dispersão Resposta Covariável.

$$L(\beta_0, \beta_1|Y) = \prod_{i=1}^{n} \frac{\exp^{\lambda_i} \lambda_i^{y}{}_i}{y_i!} \quad \text{onde} \quad \lambda_i = \exp^{\beta_0+\beta_1 x_i}$$

Em R pode-se escrever esta função

```
> verossimilhanca <- function(para, y, x) {
+     lambda <- exp(para[1] + para[2] * x)
+     l <- prod(dpois(y, lambda = lambda))
+     return(l)
+ }
```

Uma forma mais conveniente computacionalmente é escrever a log-Verossimilhança que neste caso,

```
> log.verossimilhanca <- function(para, y, x) {
+     lambda = exp(para[1] + para[2] * x)
+     ll <- sum(dpois(y, lambda = lambda, log = TRUE))
+     return(ll)
+ }
```

Como temos um modelo com dois parâmetros podemos construir um gráfico com a superfície de Verossimilhança, ou em escala de deviance para auxiliar a interpretação e construção de intervalos de confiança.

A maximização da função de Verossimilhança ou log-Verossimilhança pode facilmente ser feita numericamente.

```
> ajuste = optim(c(1, 1), log.verossimilhanca, y = y, x = x, control = list(fnscale = -1),
+     hessian = TRUE)
```

Intervalos de confiança baseados em argumentos assintóticos são facilmente obtidos.

```
> Sigma <- solve(-ajuste$hessian)
> erro <- sqrt(diag(Sigma))
> ic.b0 <- c(ajuste$par[1] - qnorm(0.975) * erro[1], ajuste$par[1],
```

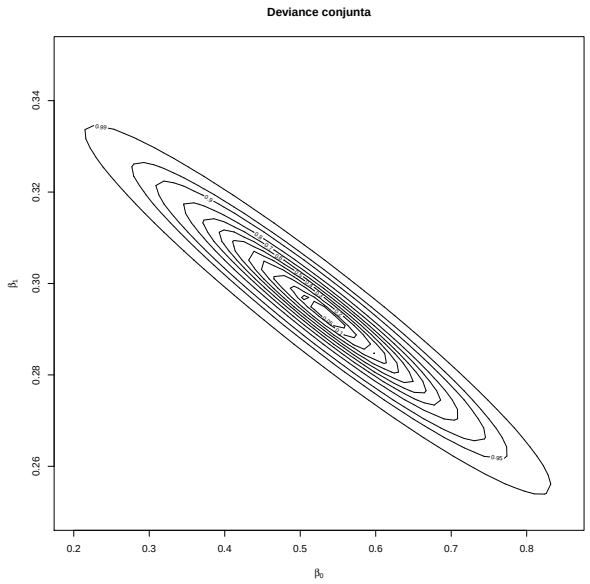

Figura 2: Deviance conjunta com limites de confiança

```
+     ajuste$par[1] + qnorm(0.975) * erro[1])
> ic.b1 <- c(ajuste$par[2] - qnorm(0.975) * erro[2], ajuste$par[2],
+     ajuste$par[2] + qnorm(0.975) * erro[2])
> cbind(ic.b0, ic.b1)
```

```
         ic.b0     ic.b1
[1,] 0.3383526 0.2666177
[2,] 0.5376172 0.2926796
[3,] 0.7368818 0.3187415
```

Estes intervalos são obtidos apartir de uma aproximação quadrática da verossimilhança, podemos sobrescrever estas funções para ver como é esta aproximação.

```
> dev.approx <- function(theta, theta.est, hessiano) {
+     de.ap <- -t(theta.est - as.numeric(theta)) %*% hessiano %*%
+         (theta.est - as.numeric(theta))
+     return(de.ap)
+ }
> deviance.app <- apply(grid, 1, dev.approx, theta.est = ajuste$par,
+     hessiano = ajuste$hessian)
```

Neste caso a aproximação quadrática tem um comportamento muito próximo da função exata. Em situações onde isso não ocorre, podemos usar um intervalo baseado em perfil de Verossimilhança. O código abaixo constrõe perfil de verossimilhança/deviance para os dois parâmetros do modelo Poisson.

```
> log.verossimilhanca <- function(b0, b1, y, x) {
+     lambda = exp(b0 + b1 * x)
```

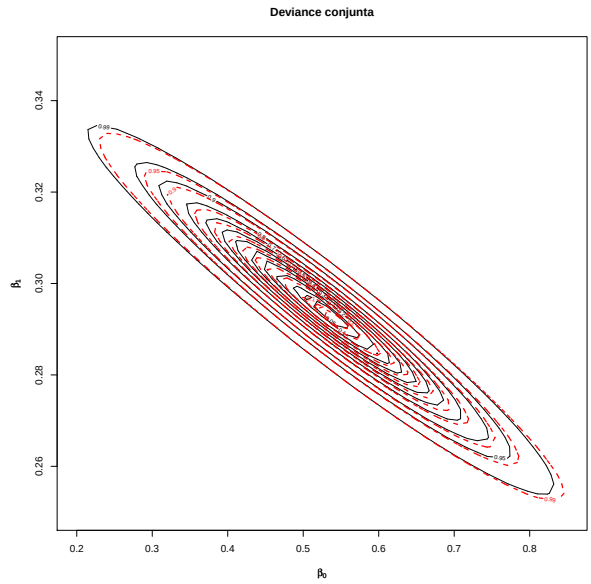

Figura 3: Deviance exata e aproximação quadrática.

```
+     ll <- sum(dpois(y, lambda = lambda, log = TRUE))
+     return(ll)
+ }
> perfil.b0 <- function(grid.b0, y, x) {
+     saida <- c()
+     perf.b0 <- function(b1, b0, y, x) {
+         log.verossimilhanca(b0 = b0, b1 = b1, y = y, x = x)
+     }
+     for (i in 1:length(grid.b0)) {
+         saida[i] <- optimize(perf.b0, interval = c(-10, 10),
+             b0 = grid.b0[i], y = y, x = x, maximum = TRUE)$objective
+     }
+     deviance.perfil <- 2 * (max(saida) - saida)
+     return(deviance.perfil)
+ }
> perfil.b1 <- function(grid.b1, y, x) {
+     saida <- c()
+     perf.b1 <- function(b0, b1, y, x) {
+         log.verossimilhanca(b0 = b0, b1 = b1, y = y, x = x)
+     }
+     for (i in 1:length(grid.b1)) {
+         saida[i] <- optimize(perf.b1, interval = c(-10, 10),
+             b1 = grid.b1[i], y = y, x = x, maximum = TRUE)$objective
+     }
+     deviance.perfil <- 2 * (max(saida) - saida)
+     return(deviance.perfil)
```

```
+ }
```

Uma vez as funções contruídas vamos usá-las

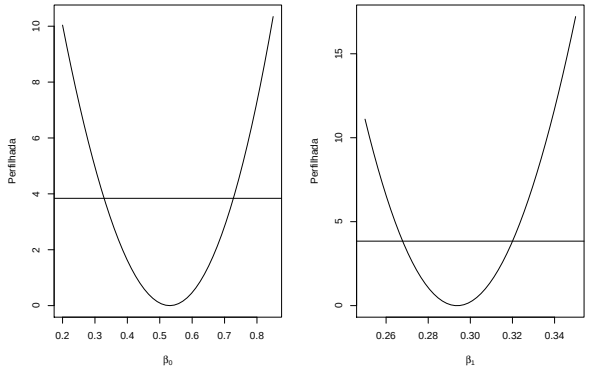

Figura 4: Perfil de verossimilhança em escala de deviance β_0 e β_1.

Para obtermos os valores realizados do intervalo de confiança baseados na verossimilhança/deviance perfilhada, devemos achar os pontos onde a deviance atinge o valor de uma distribuição Qui Quadrado com um grau de liberdade. Isto pode ser feito no R usando a função *uniroot*().

Referências

[1] DOBSON, J.; BARNETT, A.,*An Introduction to Generalized Linear Models*,New York: Chapman & Hall, 2008.