Modelos Lineares Generalizados

Testes de hipóteses

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

20 de maio de 2021

Originalmente, escrevemos o logaritmo da verossimilhança para um modelo linear generalizado como uma função $\log_e\big(L(\theta,\phi;y)\big)$ do parâmetro canônico θ para as observações. Como $\mu_i = g_c^1(\theta_i)$, para o elo canônico $g_c(\cdot)$, podemos igualmente pensar na probabilidade logarítmica como uma função da resposta esperada e, portanto, podemos escrever a verossimilhança maximizada como $\log_e\big(L(\widehat{\mu},\phi;y)\big)$.

Se dedicarmos um parâmetro a cada observação, de modo que $\widehat{\mu}_i = y_i$, por exemplo, removendo a constante do modelo de regressão e definindo um regressor fictício para cada observação, o log-verossimilhança torna-se $\log_e\big(L(y,\phi;y)\big)$.

O desvio residual sob o modelo inicial é duas vezes a diferença nessas log-verossimilhanças:

$$\begin{aligned} D(y;\widehat{\mu}) &= 2\left(\log_e\left(L(y,\phi;y)\right) - \log_e\left(L(\widehat{\mu},\phi;y)\right)\right) \\ &= 2\sum_{i=1}^{n}\left(\log_e\left(L(y_i,\phi;y_i)\right) - \log_e\left(L(\widehat{\mu},\phi;y_i)\right)\right) \\ &= 2\sum_{i=1}^{n}\frac{1}{a_i}\Big(y_i\big(g(y_i)-g(\widehat{\mu}_i)\big) - b\big(g(y_i)\big) + b\big(g(\widehat{\mu}_i)\big)\Big)\cdot \end{aligned}$$

A divisão do desvio residual pelo parâmetro de dispersão estimado produz o desvio escalonado,

$$D^*(y;\widehat{\mu}) = \frac{1}{\widetilde{\phi}}D(y;\widehat{\mu})\cdot$$

Aplicando a equação acima à distribuição Gaussiana, onde $g_c(\cdot)$ é o elo de identidade, produz $a_i = 1$ e $b(\theta) = \theta^2/2$, após alguma simplificação

$$D(y; \widehat{\mu}) = \sum_{i=1}^{n} \big(y_i - \widehat{\mu}_i\big)^2,$$

ou seja, a soma residual dos quadrados do modelo. Da mesma forma, aplicando a mesma equação à distribuição binomial, onde $g_c(\cdot)$ é a ligação logit, $a_i = n_i$ e $b(\theta) = \log_e\big(1+e^{\theta}\big)$, obtemos, após um pouco de simplificação

$$D(y; \widehat{\mu}) = \sum_{i=1}^{n} n_i \left(y_i \log_e\left(\frac{y_i}{\widehat{\mu}_i}\right) + (1 - y_i)\log_e\left(\frac{1-y_i}{1-\widehat{\mu}_i}\right)\right)\cdot$$

Estimando o parâmetro de dispersão

Observe que não exigimos uma estimativa do parâmetro de dispersão para estimar os coeficientes de regressão em um modelo linear generalizado. Embora seja, em princípio, possível estimar ϕ também por máxima verossimilhança, isso raramente é feito.

Em vez disso, lembre-se de que $\text{Var}(Y_i) = \phi a_i \nu(\mu_i)$. Resolvendo para o parâmetro de dispersão, obtemos $\phi = \text{Var}(Y_i)/a_i\nu(\mu_i)$, sugerindo o método do estimador de momentos

$$\widetilde{\phi} = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{\left(y_i - \widehat{\mu}_i\right)^2}{a_i\nu(\widehat{\mu}_i)}.$$

A matriz de covariância assintótica estimada dos coeficientes é então obtida a partir da última iteração IWLS como

$$\widehat{\text{Var}}(\widehat{\beta}) = \widetilde{\phi}(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}.$$

Como o estimador de máxima verossimilhança $\widehat{\beta}$ é normalmente distribuído assintoticamente, $\widehat{\text{Var}}(\widehat{\beta})$ pode ser usada como base para os testes Wald dos parâmetros de regressão.

Como no caso dos modelos lineares, podemos formular um teste para a hipótese linear geral

$$H_0 \ : \ \mathbf{L}_{q\times(k+1)}\beta_{(k+1)\times 1} \ = \ \mathbf{c}_{q\times 1},$$

onde a matriz de hipótese $\mathbf{L}$ e o vetor $\mathbf{c}$ do lado direito contêm constantes pré-especificadas; geralmente, $\mathbf{c} = \mathbf{0}$.

Para um modelo linear generalizado, a estatística Wald

$$Z_0^2 \ = \ (\mathbf{L}\widehat{\beta} - \mathbf{c})^\top \left(\mathbf{L}\widehat{\mathrm{Var}}(\widehat{\beta})\mathbf{L}^\top\right)^{-1}(\mathbf{L}\widehat{\beta} - \mathbf{c})$$

segue uma distribuição assintótica qui-quadrado com q graus de liberdade sob a hipótese.

A aplicação mais simples desse resultado é a estatística Wald

$$Z_0 = \widehat{\beta}_j / SE(\widehat{\beta}_j),$$

testando se um coeficiente de regressão individual é zero. Aqui, Z_0 segue uma distribuição normal padrão sob $H_0 \; : \; \beta_j = 0$ ou, equivalentemente, Z_0^2 segue uma distribuição qui-quadrado com um grau de liberdade.

Alternativamente, quando o parâmetro de dispersão é estimado a partir dos dados, podemos calcular a estatística de teste

$$F_0 \; = \; \frac{1}{q} Z_0^2,$$

que é distribuída como $F_{q,n-k-1}$ sob H_0. Aplicado a um coeficiente individual $t_0 = \pm\sqrt{F_0} = \widehat{\beta}_j / SE(\widehat{\beta}_j)$ produz um teste t com $n - k - 1$ graus de liberdade.

Ocasionalmente, é interessante testar uma hipótese ou construir um intervalo de confiança para uma função não linear dos parâmetros de um modelo linear ou linear generalizado. Se a função não linear em questão é uma função diferenciável dos coeficientes de regressão, um erro padrão assintótico aproximado pode ser obtido pelo método delta.

O método delta (Rao, 1973) emprega uma aproximação da série de Taylor de primeira ordem, isto é, linear para a função não linear. O método delta é apropriado aqui porque as estimativas de máxima verossimilhança ou quase verossimilhança dos coeficientes de um GLM são normalmente distribuídas assintoticamente.

Em pequenas amostras, no entanto, a aproximação do método delta para o erro padrão pode não ser adequada e os procedimentos de bootstrapping fornecerão resultados mais confiáveis.

Suponha que estejamos interessados na função

$$\gamma = f(\beta) = f(\beta_0, \beta_1, \cdots, \beta_k),$$

onde, por conveniência de notação, usamos β_0 para denotar a constante de regressão. A função $f(\beta)$ não precisa usar todos os coeficientes de regressão, veja o exemplo abaixo.

O estimador de máxima verossimilhança de γ é simplesmente $\widehat{\gamma} = f(\widehat{\beta})$ que, como um estimador de máxima verossimilhança, também é assintoticamente normal e a variância amostral aproximada de γ é então

$$\widehat{\text{Var}}(\widehat{\gamma}) \approx \sum_{j=0}^{k} \sum_{l=0}^{k} \widehat{\sigma}_{jk} \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_j} \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_k},$$

onde $\widehat{\sigma}_{jk}$ é o (j,k)-ésimo elemento da matriz de covariância assintótica estimada dos coeficientes, $\widehat{\text{Var}}(\widehat{\beta})$.

Para ilustrar a aplicação desse resultado, imagine que estamos interessados em determinar o valor máximo ou mínimo de uma regressão parcial quadrática. Enfocando a relação parcial entre a variável resposta e um X em particular, temos uma equação da forma

$$\mathrm{E}(Y) = \cdots + \beta_1 X + \beta_2 X^2 + \cdots.$$

A aplicação do método delta para encontrar o mínimo ou máximo de uma curva quadrática foi sugerida por Weisberg (2005).

Diferenciando esta equação em relação a X, obtemos

$$\frac{\mathrm{dE}(Y)}{\mathrm{d}X} = \beta_1 + 2\beta_2 X\cdot$$

Definindo a derivada como 0 e resolvendo para X produz o valor no qual a função atinge um mínimo, se β_2 for positivo ou um máximo se β_2 for negativo,

$$X = -\frac{\beta_1}{2\beta_2},$$

que é uma função não linear dos coeficientes de regressão β_1 e β_2.

Por exemplo, usando dados da Pesquisa Canadense de Dinâmica de Trabalho e Renda ou SLID ajustamos uma regressão de mínimos quadrados do logaritmo base 2 da taxa de salário segundo a idade quadrática, um regressor fictício para sexo e o quadrado dos anos de estudo, obtendo:

```
> SLID = read.table("https://socialsciences.mcmaster.ca/jfox/
          Books/Applied-Regression-2E/datasets/SLID-Ontario.txt",
header = T)
> head(SLID)
  age    sex compositeHourlyWages yearsEducation
1  40   Male                10.56             15
2  19   Male                11.00             13
3  46   Male                17.76             14
4  50 Female                14.00             16
5  31   Male                 8.20             15
6  30 Female                16.97             13
> ajuste = lm(I(log2(compositeHourlyWages)) ~ age+I(age^2)+
                    factor(sex)+I(yearsEducation^2), data = SLID)
```

```
> summary(ajuste)
Call:
lm(formula = I(log2(compositeHourlyWages)) ~ age + I(age^2) +
    factor(sex) + I(yearsEducation^2), data = SLID)
Residuals:
     Min       1Q   Median       3Q      Max
-3.04688 -0.34263  0.02977  0.36354  2.56370
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.725e-01  8.338e-02   6.866 7.62e-12 ***
age                  1.198e-01  4.598e-03  26.046  < 2e-16 ***
I(age^2)            -1.230e-03  5.918e-05 -20.778  < 2e-16 ***
factor(sex)Male      3.195e-01  1.796e-02  17.794  < 2e-16 ***
I(yearsEducation^2)  2.605e-03  1.135e-04  22.957  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5675 on 3992 degrees of freedom
Multiple R-squared:  0.3892,Adjusted R-squared:  0.3886
F-statistic: 635.8 on 4 and 3992 DF,  p-value: < 2.2e-16
```

Imagine que estamos interessados na idade $\gamma = \beta_1/(2\beta_2)$ em que os salários estão no máximo, mantendo o sexo e a educação constantes. As derivadas necessárias são

$$\left.\frac{d\widehat{\gamma}}{d\beta_1}\right|_{\beta=\widehat{\beta}} = -\frac{1}{2\widehat{\beta}_2} = -\frac{1}{2(-0.001230)} = 406.5$$
$$\left.\frac{d\widehat{\gamma}}{d\beta_2}\right|_{\beta=\widehat{\beta}} = -\frac{\widehat{\beta}_1}{2\widehat{\beta}_2^2} = -\frac{0.1198}{2(-0.001230)^2} = 39.593$$

Nossa estimativa pontual de γ é

$$\widehat{\gamma} = -\frac{\widehat{\beta}_1}{2\widehat{\beta}_2} = -\frac{0.1198}{2 \times 0.001230} = 48.70 \text{ anos}\cdot$$

A variância amostral estimada do coeficiente de idade é $\widehat{\text{Var}}(\widehat{\beta}_1) = 2.115 \times 10^5$, e do coeficiente de idade ao quadrado, $\widehat{\text{Var}}(\widehat{\beta}_2) = 3.502 \times 10^9$; a covariância amostral estimada para os dois coeficientes é

$$\widehat{\text{Cov}}(\widehat{\beta}_1, \widehat{\beta}_2) = 2.685 \times 10^7 \cdot$$

A variância estimada aproximada de γ é então

$$\widehat{\text{Var}}(\widehat{\gamma}) \approx 0.3419 \cdot$$

Consequentemente, o erro padrão aproximado de γ é $SE(\gamma) \approx \sqrt{0.3419} = 0.5847$, e um intervalo de confiança de aproximadamente 95% para a idade em que a renda é mais alta em média é

$$\gamma \in 48.70 \pm 1.96(0.5847) = (47.55, 49.85) \cdot$$

Vamos escrever o modelo linear generalizado em forma matricial, com preditor linear

$$\eta_{n\times 1} = \mathbf{X}_{n\times(k+1)}\beta_{(k+1)\times 1}$$

e função de ligação $g(\mu) = \eta$,onde μ é a esperança do vetor de resposta Y,

Conforme descrito anteriormente, calculamos o estimador de máxima verossimilhança $\widehat{\beta}$ de β, juntamente com o estimador da matriz de covariância assintótica $\widehat{\text{Var}}(\widehat{\beta})$ de $\widehat{\beta}$.

Considere que as linhas de $\mathbf{X}^*$ incluam regressores correspondentes a todas as combinações de valores de variáveis explicativas que aparecem em um termo de ordem superior do modelo ou, para uma variável explicativa contínua, valores abrangendo o intervalo da variável, junto com valores típicos das restantes regressores.

A estrutura de $\mathbf{X}^*$ com respeito às interações, por exemplo, é a mesma que a da matriz do modelo $\mathbf{X}$. Então os valores ajustados

$$\widehat{\eta}^* = \mathbf{X}^* \widehat{\beta},$$

representam o termo de ordem superior em questão e uma tabela ou gráfico destes valores - ou, alternativamente, dos valores ajustados transformados para a escala da variável resposta, $g^1(\widehat{\eta}^*)$ - é uma exibição de efeito.

Os erros padrão de $\widehat{\eta}^*$, disponíveis como entradas diagonais de raiz quadrada de $\mathbf{X}^* \widehat{\mathrm{Var}(\widehat{\beta})} \mathbf{X}^{*\top}$, podem ser usados para calcular intervalos de confiança pontuais para os efeitos, cujos pontos finais também podem ser transformados na escala da resposta.