CE075 - Análise de Dados Longitudinais

Silva, J.L.P.

09 de setembro, 2019

Teoria da Verossimilhança

Teoria de Verossimilhança

Considere $Y_1, Y_2, \ldots, Y_N$ respostas *iid* de uma população $f(y;\theta)$. Então a função de verossimilhança para θ é dada por

$$\mathcal{L}(\theta|y) = \prod_{i=1}^{N} f(y_i|\theta),$$

em que θ é um vetor p-dimensional de parâmetros.

O EMV (Estimador de Máxima Verossimilhança) é aquele $\hat{\theta}$ que maximiza $\mathcal{L}(\theta|y)$ ou, de forma equivalente, $l(\theta|y) = log(\mathcal{L}(\theta|y))$ no espaço de parâmetros de θ.

Teoria de Verossimilhança: Função Escore

A função escore é dada por

$$\mathcal{S}(\theta) = \frac{\partial l(\theta|y)}{\partial \theta},$$

que é p-dimensional.

O EMV é a solução do sistema de equações determinado pela função escore:

$$\mathcal{S}(\widehat{\theta}) = 0.$$

Propriedade importante: $E\left[S(\theta)\right] = 0$.

Teoria de Verossimilhança: Medida de Incerteza

$$\begin{aligned}\mathcal{I}(\theta) &= Var(\mathcal{S}(\theta)) \\ &= \mathsf{E}(\mathcal{S}(\theta)^2) \\ &= -\mathsf{E}\left(\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right),\end{aligned}$$

que é uma matriz $p \times p$ chamada de *Informação de Fisher*.

A variância assintótica de $\widehat{\theta}$ é

$$Var(\hat{\theta}) = \mathcal{I}(\theta)^{-1},$$

que é estimada avaliando θ em $\widehat{\theta}$.

Teoria de Verossimilhança: Medida de Incerteza

Usualmente é difícil encontrar o valor esperado na distribuição de Y. No entanto, podemos utilizar qualquer estimador consistente de I.

Usamos a matriz de informação observada

$$\mathcal{I}_o(\theta) = -\left(\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right),$$

que é consistente para $I(\theta)$. Ou seja,

$$Var(\widehat{\theta}) \approx \mathcal{I}_o(\theta)^{-1}.$$

Obs. O resultado é verdadeiro para qualquer estimador consistente de $\mathcal{I}$.

Estimador de Máxima Verossimilhança

Estimador de Máxima Verossimilhança

A função de verossimilhança é dada por:

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} f(y_i|\theta).$$

Dado que $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$ segue uma distribuição normal multivariada, devemos maximizar a função de log-verossimilhança:

$$l(\theta) = -\frac{K}{2} log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} log(|V_i|) - \frac{1}{2}\sum_{i=1}^{N} \{(y_i - X_i\beta)' V_i^{-1} (y_i - X_i\beta)\},$$

em que $K = \sum_{i=1}^{N} n_i$ é o número total de observações.

Observações:

- O vetor de parâmetros β somente aparece no último termo;
- Se V for fixo, o estimador de β consiste em minimizar :

$$\sum_{i=1}^{N}(y_i - X_i\beta)' V_i^{-1}(y_i - X_i\beta)$$

cuja solução é:

$$\hat{\beta}_{EMV} = (X'V^{-1}X)^{-1}X'V^{-1}Y,$$

que é conhecido como *estimador de mínimos quadrados generalizados*.

- Antes de procedermos ao caso em que precisamos estimar os componentes de variância, vejamos algumas propriedades.

Propriedades de $\hat{\beta}_{EMV}$ (assintóticas)

- $\hat{\beta}_{EMV}$ é consistente para β;
- $\hat{\beta}_{EMV}$ é assintoticamente normal (Wald):

$$\sqrt{K}(\hat{\beta}_{EMV} - \beta) \xrightarrow{\mathcal{D}} N(0, \mathcal{I}^{-1})$$

- Estas propriedades valem assintoticamente, mesmo se y_i não tiver distribuição normal multivariada (para dados completos).
- A distribuição amostral de $\hat{\beta}$ quando V_i é estimada dos dados é aproximadamente normal multivariada com média β e covariância:

$$Cov(\hat{\beta}) = \left\{ \sum_{i=1}^{N} (X_i' V_i^{-1} X) \right\}^{-1}.$$

Inferência

Para a construção de intervalos de confiança e testes de hipóteses para β, podemos fazer uso direto da estimativa $\hat{\beta}$ e sua matriz de covariância estimada

$$\widehat{Cov}(\hat{\beta}) = \left\{ \sum_{i=1}^{N} (X_i' \hat{V}_i^{-1} X) \right\}^{-1},$$

em que V_i é substituído por $\hat{V}_i$, o EMV de V.

Intervalos de confiança podem ser construídos para qualquer componente de β. Por exemplo, um IC de 95% de confiança para β_k é dado por:

$$\hat{\beta}_k \pm 1.96 \sqrt{\widehat{Var}(\beta_k)}.$$

Inferência

Similarmente, um teste para a hipótese nula $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0$ pode ser baseado na seguinte estatística Wald:

$$Z = \frac{\hat{\beta}_k}{\sqrt{\widehat{Var}(\beta_k)}},$$

cujo valor deve ser comparado com uma distribuição normal padrão.

De forma mais geral, pode ser de interesse construir intervalos de confiança e testar hipóteses sobre combinações lineares dos componentes de β.

Inferência

Seja L um vetor ou matriz de pesos conhecidos e suponha que seja de interesse testar a hipótese $H_0 : L\beta = 0$. A combinação $L\beta$ representa um contraste de interesse científico.

Por exemplo, suponha que $\beta = (\beta_1, \beta_2, \beta_3)'$:

- se $L = (0, 0, 1)$ então $H_0 : L\beta = 0$ é equivalente a $H_0 : \beta_3 = 0$;
- se $L = (0, 1, -1)$ então $H_0 : L\beta = 0$ é equivalente a $H_0 : \beta_2 - \beta_3 = 0$ ou $H_0 : \beta_2 = \beta_3$.

A distribuição amostral de $L\hat{\beta}$ é multivariada normal com média $L\beta$ e matriz de covariância $LCov(\hat{\beta})L'$.

Inferência

Quando L é escalar, um IC aproximado de 95% para $L\beta$ é dado por

$$L\hat{\beta} \pm 1.96\sqrt{L\widehat{Cov}(\beta)L'}.$$

Similarmente, podemos testar $H_0 : L\beta = 0$ vs $H_1 : L\beta \neq 0$ por meio da estatística Wald:

$$Z = \frac{L\hat{\beta}}{\sqrt{L\widehat{Cov}(\beta)L'}}.$$

Considere agora o caso em que L tem mais de uma linha. Por exemplo, suponha que $\beta = (\beta_1, \beta_2, \beta_3)'$ e que seja de interesse testar a igualdade dos três parâmetros de regressão.

Inferência

A hipótese nula é $H_0 : \beta_1 = \beta_2 = \beta_3$. Fazendo

$$L = \begin{pmatrix} -1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix},$$

temos $H_0 : L\beta = 0$, pois

$$L\beta = \begin{pmatrix} -1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 - \beta_2 \\ \beta_1 - \beta_3 \end{pmatrix} = 0,$$

então

$$\begin{pmatrix} \beta_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix},$$

ou, equivalentemente, $\beta_1 = \beta_2 = \beta_3$.

Inferência: Teste Wald Multivariado

Em geral, se L tem r linhas, então um teste simultâneo de $H_0 : L\beta = 0$ é dado por

$$W = (L\hat{\beta})' \left\{ L\widehat{Cov}(\hat{\beta})L' \right\}^{-1} (L\hat{\beta}),$$

que segue uma distribuição χ^2 com r graus de liberdade.

Uma alternativa ao teste Wald é o *Teste da Razão de Verossimilhanças*, que compara a verossimilhança de dois modelos: um que incorpora a restrição $L\beta = 0$ (modelo reduzido) e o outro irrestrito (modelo completo).

Inferência: Teste da Razão de Verossimilhanças

O teste formal é obtido tomando-se duas vezes a diferença nas log-verossimilhanças maximizadas:

$$G = 2 \times (\hat{l}_{completo} - \hat{l}_{reduzido}),$$

cujo valor é comparado com uma distribuição χ^2 com graus de liberdade igual a diferença entre o número de parâmetros nos modelos completo e reduzido.

Também é possível a construção de limites de confiança para β ou $L\beta$ baseados na *verossimilhança perfilada*.

Comentários

- As distribuições de referência (assintótica) normal e qui-quadrado são utilizadas como aproximações da t e da F, respectivamente. É possível estimar os gl para utilizar a t e a F, especialmente para amostras de tamanho pequeno.
- O valor-p obtido através da estatística de Wald é menor do que o verdadeiro (e será tão menor quanto menor for o tamanho da amostra).
- Devemos evitar o uso da estatística de Wald para testar os componentes de variância pois a convergência para normal é lenta para amostras pequenas e variâncias próximas de zero.
- Desta forma, o recomendado é a estatística da razão de verossimilhança.

Estimador de Máxima Verossimilhança Restrita

Discutimos a inferência sobre β e os componentes de variância com base na maximização da função de log-verossimilhança:

$$l(\theta) = -\frac{K}{2} log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} log(|V_i|) - \frac{1}{2}\sum_{i=1}^{N}\Big\{(y_i - X_i\beta)' V_i^{-1}(y_i - X_i\beta)\Big\},$$

Embora as estimativas de máxima verossimilhança de β e V_i tenham propriedades assintóticas desejáveis, é sabido que possuem viés de amostras pequenas. Por exemplo, os elementos diagonais de V_i são subestimados.

Para ilustrar, considere o caso de um estudo transversal.

Estimador de Máxima Verossimilhança Restrita (ou Residual)

Estudo linear-normal transversal EMV (amostra de tamanho N)

$$\hat{\sigma}^2 = \frac{SQR}{N} = \frac{(y - x\hat{\beta})'(y - x\hat{\beta})}{N}$$

$$E[\hat{\sigma}^2] = \frac{N}{N - p}\sigma^2$$

- **Razão:** EMV não leva em consideração que β é estimado pelos dados;
- **Proposta:** utilizar EMVR (Estimador Máxima Verossimilhança Restrita);
- **Ideia:** Separar as partes dos dados para estimar α daqueles utilizados para estimar β.

Estimador de Máxima Verossimilhança Restrita (ou Residual)

Transformar a resposta Y tal que a distribuição resultante não dependa de β. Ou seja,

$$Z = AY \quad \text{tal que} \quad E(Z) = 0$$

Exemplo: Modelo Linear-Normal Transversal

$$A = I - H = I - X(X'X)^{-1}X'$$

$$E(Z) = (I - X(X'X)^{-1}X')E(Y) = X\beta - X(X'X)^{-1}X'X\beta = 0$$

Transformação $y \rightarrow (Z, \hat{\beta})$

A função de log-verossimilhança para Z escrita em termos de Y e $\hat{\beta}$ é

$$l^*(\alpha) = -\frac{1}{2}\sum_{i=1}^{N} log|V_i| - \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i\hat{\beta})' V_i^{-1}(y_i - x_i\hat{\beta}) - \frac{1}{2}\left|\sum_{i=1}^{N} x_i' V_i^{-1} x_i\right|$$

Na log-verossimilhança residual acima foi feita uma correção para o fato de que β também foi estimado.

Observação

O termo adicional da função de log-verossimilhança restrita é:

$$-\frac{1}{2} log \left| \sum_{i=1}^{N} x_i' V_i^{-1} x_i \right| = \frac{1}{2} log \left| \left(\sum_{i=1}^{N} x_i' V_i^{-1} x_i \right)^{-1} \right| = log |Cov(\widehat{\beta})|^{1/2}.$$

- A verossimilhança foi multiplicada por um fator que é raiz quadrada da variância generalizada de $\hat{\beta}$, uma medida resumo da variação da estimativa de β.
- Este termo é o equivalente a fazer a correção no denominador de $\hat{\sigma}^2$.

Processo de Estimação: EMVR

1. Estimar β por:

$$\begin{aligned}\hat{\beta} &= (X'V^{-1}X)^{-1}XV^{-1}Y \\ &= \left(\sum_{i=1}^{N} x_i' V_i^{-1} x_i\right)^{-1} \left(\sum_{i=1}^{N} x_i' V_i^{-1} y_i\right);\end{aligned}$$

2. Encontrar o EMVR para α a partir de $l^*(\alpha)$;
3. Continuar este processo até a convergência.

Processo de Estimação: EMVR

- O EMVR é recomendado para α quando comparado ao EMV. No entanto, a correção do vício se torna *desprezível* quando Nn é muito maior que p;
- A estatística da Razão de MVR pode ser usado para comparar modelos de covariâncias aninhadas mas não pode ser utilizado para comparar modelos aninhados para a média.
- O termo extra do determinante depende da especificação do modelo de regressão; assim verossimilhanças para dois modelos aninhados para a média são baseados em dois conjuntos completamente diferentes de respostas transformadas, o que tornaria a comparação sem sentido.
- Modelos de covariância não aninhados podem ser comparados via estatísticas AIC ou BIC.

Modelos Marginais

Modelos Marginais: GEE/EMVR

- GEE e EMVR são similares (igualmente eficientes) com dados completos.
- A única condição para GEE produzir inferências válidas é a estrutura da média estar corretamente especificada.
- Especificando corretamente a estrutura de variância-covariância ganha-se em eficiência no processo inferencial.
- Na presença de dados faltantes (MAR e NMAR), o GEE não produz inferências válidas. Por outro lado, o EMVR produz inferências válidas nesta condição (somente MAR) se a distribuição normal for corretamente especificada para a resposta.