Regressão Múltipla

Parte II:Intervalos de confiança e testes de hipóteses para os parâmetros, soma extra de quadrados do modelo e dados padronizados.

Inferência sobre os parâmetros da regressão

Os estimadores de mínimos quadrados ou de máxima verossimilhança são não tendenciosos, isto é: $E(\hat{\beta}) = \beta$

A matriz de variância-covariância dos estimadores é dada por:

$$\underset{(p\ x\ p)}{\boldsymbol{\sigma}^2(\hat{\beta})} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \qquad \textit{(28)}$$

A estimativa da matriz de variância-covariância é dada por:

$$\underset{(p\ x\ p)}{s^2(\hat{\boldsymbol{\beta}})} = MS_R(\mathbf{X}'\mathbf{X})^{-1} \quad (29)$$

Intervalo de confiança para os parâmetros β_k

Para o modelo com erros normais, (17), temos:

$$\frac{\hat{\beta}_k - \beta_k}{s(\hat{\beta}_k)} \sim t(n-p) \qquad k = 0,1,\ldots,p-1 \qquad (30)$$

Assim, o intervalo para β_k, com confiança $1-\alpha$ é dado por:

$$\hat{\beta}_k \pm t(1-\alpha/2; n-p)s(\hat{\beta}_k) \qquad (31)$$

Testes de hipóteses para β_k

Hipóteses:

$$H_0 : \beta_k = 0$$
$$H_a : \beta_k \neq 0 \qquad (32)$$

Estatística de teste:

$$t^* = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \qquad (33)$$

Critério do teste:

Se $|t^*| \leq t(1-\alpha/2;n-p)$, aceita-se a hipótese nula, caso contrário rejeita-se a mesma.

Estimação da resposta média e predição de uma nova observação

Intervalo de confiança para $E(Y_h)$

Para valores dados de $X_1, X_2, \ldots, X_{P-1}$, representados por: $X_{h1}, X_{h2}, \ldots, X_{h,P-1}$, a resposta média é representada por $E(Y_h)$. Vamos definir o vetor:

$$\underset{p\,x\,1}{\mathbf{X_h}} = \begin{bmatrix} 1 \\ X_{h1} \\ . \\ . \\ X_{h,p-1} \end{bmatrix}$$

A resposta média estimada, correspondente ao vetor $\mathbf{X}_h$, é dada por :

$$\hat{Y}_h = \mathbf{X}_h^{'}\hat{\boldsymbol{\beta}} \qquad (34)$$

Estimação da resposta média e predição de uma nova observação

A variância estimada da resposta média é dada por:

$$s^2(\hat{Y}_h) = MS_R(\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h) = \mathbf{X}_h'\mathbf{s}^2(\hat{\beta})\mathbf{X}_h \qquad (35)$$

O intervalo de confiança para a resposta média, $E(Y_h)$, é dado por:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s(\hat{Y}_h) \qquad (36)$$

Limites de predição para uma nova observação $Y_{h(novo)}$

Os limites de predição com confiança *1-*α para uma nova observação $Y_{h(nova)}$ correspondente ao vetor $\mathbf{X}_h$, os valores das variáveis explanatórias, são:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s(pred) \quad (37)$$

A variância do erro de predição (é a diferença entre a nova observação e o valor estimado) é dado por:

$$s^2(pred) = MS_R(1+\mathbf{X}_h^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}_h) \quad \textbf{(38)}$$

Método de soma extra de quadrados

- A idéia básica é verificar a redução na soma de quadrados do erro quando uma ou mais variáveis preditoras são adicionadas no modelo de regressão, dado que outras variáveis preditoras já estão incluídas no modelo. De outro lado, podemos pensar no acréscimo na soma de quadrados da regressão quando uma ou mais variáveis explanatórias são adicionadas no modelo.
- Utilização: verificar se certas variáveis *X* podem ser retiradas do modelo de regressão. (Construção de modelos).

Método de soma extra de quadrados

- Consider o modelo com p-1 preditores.
- Seja o vetor de coeficientes particionados como:
- Considere as hipóteses

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

com p-r colunas

com r colunas

$$H_0 : \boldsymbol{\beta}_2 = 0$$

$$H_a : \boldsymbol{\beta}_2 \neq 0$$

Método de soma extra de quadrados

- Modelo completo $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$
- $\mathbf{X}_1$ n×(p-r) – representa as colunas dos valores de X associado com $\boldsymbol{\beta}_1$
- $\mathbf{X}_2$ n×r – representa coluna dos valores de X associado com $\boldsymbol{\beta}_2$
- Para o modelo completo $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ temos

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}^T y$$

- Soma de Quadrados do modelo completo com p-1 graus de liberdade e soma de quadrados dos resíduos com n-p graus de liberdade

$$SS_M = \hat{\boldsymbol{\beta}} X^T y \longrightarrow MS_R = \frac{\boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}} X^T y}{n - p}$$

Método de soma extra de quadrados

- Para encontrar a contribuição de $\boldsymbol{\beta}_2$ no modelo, ajuste o modelo assumindo que $\boldsymbol{\beta}_2=\mathbf{0}$ (hipótese nula é verdadeira).
- O modelo reduzido é $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$
- Encontra $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1\mathbf{X}_1^T)^{-1}\mathbf{X}_1^T y$
- A soma de quadrados da regressão com p-r graus de liberdade é

$$SS_M(\boldsymbol{\beta}_1) = \hat{\boldsymbol{\beta}}_1 \boldsymbol{X}_1^T y$$

- A soma de quadrados devido a $\boldsymbol{\beta}_2$ dado $\boldsymbol{\beta}_1$ com p-(p-r)=r graus de liberdade

$$SS_M(\boldsymbol{\beta}_2 / \boldsymbol{\beta}_1) = SS_M(\boldsymbol{\beta}) - SS_M(\boldsymbol{\beta}_1)$$

Soma extra de quadrados

Método de soma extra de quadrados

- Estatística do teste

$$F^* = \frac{SS_M(\beta_2 / \beta_1) / r}{MS_R}$$

Se $F^* > F(\alpha; r, n-p)$, rejeitamos a hipótese nula concluindo que pelo menos um dos parâmetros in $\boldsymbol{\beta}_2$ é diferente de zero. Caso contrário, não rejeitamos a hipótese nula.

A soma extra de quadrados visa encontrar o melhor subconjunto de regressores para o problema.

Caso especial: Se as colunas de $\mathbf{X}_1$ são ortogonais as colunas de $\mathbf{X}_2$ ($\mathbf{X}_2^T \mathbf{X}_1 = \mathbf{0}$)

- Esse caso permite encontrar a soma de quadrados da regressão devido somente a β_2 ou seja $SS_M(\beta_2)$ sem levar em conta a dependência no regressor $\mathbf{x}_1$.

$$\begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1^T y \\ X_2^T y \end{bmatrix}$$

- Nessas condições a solução das equações normais são

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \mathbf{X}_1^T y \qquad \hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \mathbf{X}_2^T y$$

Caso especial: Se as colunas de $\mathbf{X}_1$ são ortogonais as colunas de $\mathbf{X}_2$ ($\mathbf{X}_2^T \mathbf{X}_1 = \mathbf{0}$)

- A soma de quadrados da regressão para o modelo completo é

$$SS_M(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}\boldsymbol{X}^T y = \hat{\boldsymbol{\beta}}_1 \boldsymbol{X}_1^T y + \hat{\boldsymbol{\beta}}_2 \boldsymbol{X}_2^T y$$

- Para cada conjunto temos

$$SS_M(\boldsymbol{\beta}_1) = \hat{\boldsymbol{\beta}}_1 \boldsymbol{X}_1^T y \qquad SS_M(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}_2 \boldsymbol{X}_2^T y$$

- Portanto,

$$SS_M(\boldsymbol{\beta}_1 / \boldsymbol{\beta}_2) = SS_M(\boldsymbol{\beta}) - SS_M(\boldsymbol{\beta}_2) \equiv SS_M(\boldsymbol{\beta}_1)$$

$$SS_M(\boldsymbol{\beta}_2 / \boldsymbol{\beta}_1) = SS_M(\boldsymbol{\beta}) - SS_M(\boldsymbol{\beta}_1) \equiv SS_M(\boldsymbol{\beta}_2)$$

Padronização dos dados

- É útil quando os regressores estão em escalas de unidades diferentes.
- Dois tipos mais comuns:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j} \qquad y_i^* = \frac{y_i - \bar{y}}{S_y}$$

i=1,2,…,n

j=1,2,…,p-1

$$y_i^* = \beta_0 + \beta_1 z_{i1} + \ldots + \beta_p z_{ip-1} + \varepsilon_i$$

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}} \qquad y_i^* = \frac{y_i - \bar{y}}{\sqrt{SS_T}}$$

j=1,2,…,p-1

i=1,2,…,n

$$y_i^* = \beta_0 + \beta_1 w_{i1} + \ldots + \beta_p w_{ip-1} + \varepsilon_i$$

Cada regressor tem média zero e $\sqrt{\sum_{i=1}^{n} (w_{ij} - \bar{w}_j)^2} = 1$