DSBD - Modelos Lineares

Slides: Cesar Taconeli, Apres.: José Padilha

11 de agosto, 2018

Aula 3 - Regressão linear múltipla

Introdução

- A regressão linear múltipla pode ser vista como uma extensão da regressão linear simples, em que **um conjunto de variáveis independentes** são utilizadas para explicar a resposta.

- Ao considerar conjuntamente o efeito de duas ou mais variáveis independentes temos condições de avaliar o efeito de uma particular variável ajustado (controlando) o efeito das demais variáveis.

- A análise de regressão linear com múltiplas variáveis é mais complexa que a regressão linear simples devido à maior dificuldade de encontrar um bom modelo, visualizar e interpretar os resultados.

Definição e propriedades do modelo

- O modelo de regressão linear múltipla é definido da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i. \quad (1)$$

- Para os erros do modelo de regressão linear múltipla assumimos suposições semelhantes às consideradas para o modelo de regressão linear simples:
  - Linearidade: $E(\epsilon_i) = 0$;
  - Variância constante: $Var(\epsilon_i) = \sigma^2$;
  - Independência: ϵ_i e ϵ_j são independentes para $i \neq j$;
  - x_i é independente de ϵ_i, para todo i;
  - Normalidade: $\epsilon_i \sim N(0, \sigma^2)$.

Definição e propriedades do modelo

- Como consequências da especificação do modelo, temos:

1. $E(y_i|\boldsymbol{x_i} = (x_{i1}, x_{i2}, ..., x_{ik})') = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$;
2. $Var(y_i|\boldsymbol{x_i}) = \sigma^2$;
3. $y_i|\boldsymbol{x_i} \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik},\ \sigma^2)$;
4. Condicional aos respectivos vetores de variáveis explicativas, y_i e y_j são independentes, para todo $i \neq j$.

- Observe que:

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} = \beta_j \qquad (2)$$

Interpretação dos parâmetros do modelo de regressão linear múltipla

β_j representa a alteração esperada na resposta (y) para uma unidade a mais em x_j quando todas as demais variáveis $x_k \neq x_j$ são mantidas fixas.

- Devido a essa interpretação, os parâmetros de regressão $(\beta_j' s)$ são usualmente chamados **coeficientes de regressão parcial**.

Definição e propriedades do modelo

- O intercepto (β_0) é a resposta esperada no ponto $x_1 = 0$, $x_2 = 0$, ..., $x_k = 0$, caso esse ponto pertença ao escopo do problema;

- A interpretação apresentada para os parâmetros $\beta_j' s$ somente é válida na ausência de interações;

- Considere o seguinte modelo de regressão linear múltipla com termo de interação:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon. \tag{3}$$

Definição e propriedades do modelo

- Nesse caso, por exemplo:

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x_1} = \beta_1 + \beta_3 x_2. \tag{4}$$

- Assim, mantendo x_2 fixa, espera-se uma variação de $\beta_1 + \beta_3 x_2$ em y para cada unidade acrescida em x_1.
- De forma semelhante, mantendo x_1 fixa, espera-se uma variação de $\beta_2 + \beta_3 x_1$ em y para cada unidade acrescida em x_2.
- Dessa forma, a superfície de regressão não é mais plana, pois a taxa de variação de x_1 varia conforme o valor de x_2 e vice-versa.

Notação matricial do modelo

- Considere n observações $(y_i, \boldsymbol{x}_i)$, em que $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ik})'$:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_k x_{1k} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{2k} + \epsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_k x_{nk} + \epsilon_n$$

Notação matricial do modelo

- O modelo de regressão linear múltipla fica representado matricialmente por:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon, \tag{5}$$

em que

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Notação matricial do modelo

- As especificações e resultados apresentados anteriormente podem ser representados na forma matricial:

1. $E(\boldsymbol{\epsilon}) = \mathbf{0}$;
2. $Var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$;
3. $E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\beta$;
4. $Var(\boldsymbol{y}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}$;

22. $\boldsymbol{y}|\boldsymbol{X} \sim Normal(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I})$.

Estimação dos parâmetros do modelo por mínimos quadrados

- Vamos considerar n observações $(y_i, \boldsymbol{x}_i)$, em que $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ik})'$.
- A estimação de mínimos quadrados para o modelo de regressão linear múltipla baseia-se, novamente, na determinação de $\beta_0, \beta_1, ..., \beta_k$ que minimizem a soma de quadrados dos erros:

$$S = S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2 \quad (6)$$

Estimação dos parâmetros do modelo por mínimos quadrados

- Os estimadores de mínimos quadrados para $\beta_0, \beta_1, ..., \beta_k$ devem satisfazer:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_0} \\ \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_k} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{7}$$

Estimação dos parâmetros do modelo por mínimos quadrados

- Derivando parcialmente em relação aos parâmetros de regressão obtemos:

$$\frac{\partial S}{\partial \beta_0}\bigg|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right) = 0 \tag{8}$$

e

$$\frac{\partial S}{\partial \beta_j}\bigg|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right) x_{ij} = 0, \quad j = 1, 2, \ldots, k. \tag{9}$$

Estimação dos parâmetros do modelo por mínimos quadrados

- Na forma matricial:

$$S(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta), \tag{10}$$

de maneira que o vetor $\hat{\beta}$ tal que:

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = \mathbf{0} \tag{11}$$

é o estimador de mínimos quadrados de β, dado por:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{12}$$

Estimação dos parâmetros do modelo por mínimos quadrados

- Observe que os estimadores de mínimos quadrados somente existem se a matriz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ existe;

- A condição de existência de $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ é que as colunas de $\boldsymbol{X}$ sejam linearmente independentes, ou seja, que nenhuma coluna de $\boldsymbol{X}$ seja combinação linear das demais;

- O modelo ajustado, para um vetor $\boldsymbol{x} = (1, x_1, x_2, ..., x_k)$ fica denotado por:

$$\hat{y} = \boldsymbol{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k \tag{13}$$

Estimação dos parâmetros do modelo por mínimos quadrados

- O vetor de valores ajustados $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$ é dado por:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}. \tag{14}$$

- A matriz $n \times n$ $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, chamada matriz chapéu (*hat matrix*) mapeia o vetor de valores observados no vetor de valores ajustados.
- O vetor de resíduos $\boldsymbol{r} = (r_1, r_2, ..., r_n)$ fica definido, em notação matricial, por:

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}, \tag{15}$$

em que $r_i = y_i - \hat{y}_i$, $i = 1, 2, ..., n$.

Estimação dos parâmetros do modelo por mínimos quadrados

- Propriedades dos estimadores:

1. $E(\hat{\beta}) = \beta$ ($\hat{\beta}$ é um estimador não viciado de β);

2. $Var(\hat{\beta}) = \sigma^2(\boldsymbol{X'X})^{-1}$;

3. $\hat{\beta}$ é o melhor (mais eficiente) estimador linear não viciado de β (teorema de Gauss Markov);

4. Sob a suposição de que os erros têm distribuição normal os estimadores de mínimos quadrados equivalem aos de máxima verossimilhança.

Estimação de σ^2

- Um estimador não viciado para σ^2, baseado na soma de quadrados de resíduos, é dado por:

$$\hat{\sigma}^2 = QM_{Res} = \frac{SQ_{Res}}{n-p} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}, \tag{16}$$

em que $p = k + 1$ é o número de parâmetros do modelo.

Testes de hipóteses e intervalos de confiança para os parâmetros do modelo de RLM

- Assim como no caso de RLS, também na RLM a inferência sobre os parâmetros do modelo é um ponto importante, que permitirá:

```
* Checar a significância do modelo ajustado;

* Identificar quais variáveis explicativas são relevantes na a

* Avaliar o erro de estimativas e predições geradas pelo model
```

- Deste ponto em diante assumiremos todas as suposições especificadas para os erros, inclusive a de normalidade.

Análise de variância

- Na análise de variância em regressão linear múltipla, a variação total (corrigida pela média) é novamente decomposta em dois componentes: variação explicada pela regressão e variação residual, tal que:

$$SQ_{Total} = SQ_{Reg} + SQ_{Res}. \tag{17}$$

- Usando notação matricial, as somas de quadrados ficam definidas por:

$$SQ_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}; \tag{18}$$

$$SQ_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}; \tag{19}$$

$$SQ_{Total} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \boldsymbol{y}'\boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}. \tag{20}$$

Análise de variância

Tabela 1: Quadro de análise de variânciapara o modelo de RLM

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F
Regressão	$p-1$	$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$	$QM_{Reg} = \frac{SQ_{Reg}}{p-1}$	$F = \frac{QM_{Reg}}{QM_{Res}}$
Resíduos	$n-p$	$\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$	$QM_{Res} = \frac{SQ_{Res}}{n-p}$	
Total	$n-1$	$\mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$		

- Vale lembrar que n é o tamanho da amostra e $p = k + 1$ o número de parâmetros do modelo.

Análise de variância

- Podemos testar a significância do modelo ajustado com base no seguinte par de hipóteses:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0;$$

$$H_1 : \beta_j \neq 0 \text{ para pelo menos um } j.$$

- Sob a hipótese nula (não significância do modelo) a estatística F segue distribuição $F-$Snedecor, com $p-1$ e $n-p$ graus de liberdade.
- Assim, fixado um nível de significância α, H_0 deve ser rejeitada se o valor da estatística F for maior que o quantil $1-\alpha$ da distribuição $F_{p-1,n-p}$.

Análise de variância

- O coeficiente de determinação, como anteriormente, fica definido por:

$$R^2 = 1 - \frac{SQ_{Res}}{SQ_{Total}} = \frac{SQ_{Reg}}{SQ_{Total}}, \qquad (21)$$

que expressa a proporção da variabilidade original dos dados explicada pelo modelo de regressão.

- Uma propriedade de R^2 que o torna pouco apropriado para a comparação dos ajustes de diferentes modelos é que ele nuna decresce à medida que incluímos novas variáveis ao modelo.

Análise de variância

- Como alternativa ao R^2 podemos considerar o R^2 ajustado, definido por:

$$R^2_{Aj} = 1 - \frac{SQ_{Res}/(n-p)}{SQ_{Total}/(n-1)}. \tag{22}$$

- Como $SQ_{Total}/(n-1)$ é fixo, então R^2_{Aj} somente aumentará se houver redução do quadrado médio de resíduos.
- Diferentemente de R^2, R^2_{Aj} penaliza a inclusão de variáveis não importantes no modelo, permitindo comparar adequadamente modelos com diferentes complexidades (números de variáveis).

TH's e IC's para os parâmetros do modelo de regressão linear múltipla

- Primeiramente vamos considerar TH's e IC's para parâmetros individuais do modelo.
- Suponha que se deseja testar a significância de x_j no modelo. Partimos do seguinte par de hipóteses:

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0. \tag{23}$$

- A estatística do teste é dada por:

$$t = \frac{\hat{\beta}_j}{ep(\hat{\beta}_j)}, \tag{24}$$

em que $ep(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 (\boldsymbol{X'X})^{-1}_{jj}}$, sendo $(\boldsymbol{X'X})^{-1}_{jj}$ o $j-ésimo$ termo da diagonal de $(\boldsymbol{X'X})^{-1}$ e $\hat{\sigma}^2 = QM_{Res}$.

TH's e IC's para os parâmetros do modelo de regressão linear múltipla

- Sob a hipótese nula a estatística t tem distribuição $t - Student$ com $n - p$ graus de liberdade.
- Assim, a hipótese H_0 deverá ser rejeitada, para um nível de significância α, se $|t| > |t_{n-p,\alpha/2}|$, em que $t_{n-p,\alpha/2}$ é o quantil $\alpha/2$ da distribuição $t - Student$ com $n - p$ graus de liberdade.
- Usando a distribuição t_{n-p} como referência, um intervalo de confiança $100(1 - \alpha)\%$ para β_j fica definido por:

$$\hat{\beta}_j \mp t_{n-p,\alpha/2}\sqrt{\hat{\sigma}^2(\boldsymbol{X'X})^{-1}_{jj}}. \tag{25}$$

- Para qualquer valor β_{j0} pertencente ao intervalo de confiança não se tem evidências, ao nível de significância α, que $\beta_j \neq \beta_{j0}$.

Intervalo de confiança para a resposta média e para uma predição

- Considere interesse em estimar a resposta média em um ponto $\boldsymbol{x'}_0 = (1, x_{01}, x_{02}, ..., x_{0k})$, ou seja, $E(y|\boldsymbol{x}_0)$.
- A estimativa pontual é dada pelo valor ajustado pelo modelo em $\boldsymbol{x}_0$:

$$\widehat{E(y|\boldsymbol{x}_0)} = \hat{y}_0 = \boldsymbol{x'_0}\hat{\beta}. \tag{26}$$

- O estimador apresentado é não viciado para a real resposta média, com variância:

$$Var(\widehat{E(y|\boldsymbol{x}_0)}) = \sigma^2 \boldsymbol{x'_0}(\boldsymbol{X'X})^{-1}\boldsymbol{x_0}. \tag{27}$$

Intervalo de confiança para a resposta média e para uma predição

- Um intervalo de confiança $100(1-\alpha)\%$ para a resposta média em $\boldsymbol{x'}_0 = (1, x_{01}, x_{02}, ..., x_{0k})$ é dado por:

$$\widehat{E(y|\boldsymbol{x}_0)} \mp t_{n-p,\alpha/2}\sqrt{\hat{\sigma}^2 \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}}. \tag{28}$$

em que $\hat{\sigma}^2 = QM_{Res}$.

- Considere agora que se deseja predizer a resposta em um ponto $\boldsymbol{x'}_0 = (1, x_{01}, x_{02}, ..., x_{0k})$.
- A estimativa pontual, novamente, é dada pelo valor ajustado de y em $\boldsymbol{x'}_0$:

$$\hat{y}_0 = \mathbf{x_0'}\hat{\boldsymbol{\beta}}. \tag{29}$$

Intervalo de confiança para a resposta média e para uma predição

- Neste caso, a variância de $\hat{y}_0$ fica dada por:

$$Var(\hat{y}_0) = \sigma^2 \left(1 + \boldsymbol{x_0'}(\boldsymbol{X'X})^{-1}\boldsymbol{x_0}\right). \tag{30}$$

- Um intervalo de confiança $100(1-\alpha)\%$ para a predição de uma nova observação em $\boldsymbol{x}_0$ fica dada por:

$$\hat{y}_0 \mp t_{n-p,\alpha/2}\sqrt{\hat{\sigma}^2 \left(1 + \boldsymbol{x_0'}(\boldsymbol{X'X})^{-1}\boldsymbol{x_0}\right)}, \tag{31}$$

em que $\hat{\sigma}^2 = QM_{Res}$.

Inferência para vários parâmetros do modelo

- Em geral os estimadores dos parâmetros do modelo de RLM são correlacionados (a menos que as correspondentes variáveis sejam independentes);

- Avaliar a significância individual das variáveis e avaliar a significância conjunta das mesmas, neste caso, são coisas distintas.

- Em alguns casos temos interesse particular em analisar a significância conjunta de dois ou mais parâmetros, como no caso de modelos polinomiais e com variáveis indicadoras.

- Vamos abordar testes de hipóteses simultâneos para dois ou mais parâmetros do modelo usando o princípio da verossimilhança.

Inferência para vários parâmetros do modelo

- Considere o modelo de regressão linear múltipla:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon, \quad (32)$$

$\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ o vetor de estimativas de mínimos quadrados e $\hat{\sigma}^2 = SQ_{Res}/n$ a estimativa de máxima verossimilhança para σ^2.

- O interesse aqui é testar uma hipótese do tipo $H_0 = \beta_1 = \beta_2 = ... = \beta_q = 0$, $q \leq k$. Por simplicidade de notação, vamos considerar que a hipótese nula contemple os q primeiros parâmetros do modelo.

Inferência para vários parâmetros do modelo

- O modelo induzido pela hipótese nula é dado por:

$$\begin{aligned} y &= \beta_0 + 0x_1 + 0x_2 + ... + 0x_q + \beta_{q+1}x_{q+1} + ... + \beta_k x_k + \epsilon \\ &= \beta_0 + \beta_{q+1}x_{q+1} + ... + \beta_k x_k + \epsilon \end{aligned} \tag{33}$$

- Vamos denotar por $\hat{\boldsymbol{\beta}}'_{\mathbf{0}} = (\hat{\beta}'_0, 0, 0, ..., \hat{\beta}'{}_{q+1}, ..., \hat{\beta}'{}_k)$ o estimador de mínimos quadrados para o modelo restrito.

Inferência para vários parâmetros do modelo

- A verossimilhança para o modelo completo, avaliada nas estimativas de máxima verossimilhança, é dada por:

$$L = \left(2\pi \frac{SQ_{Res}}{n}\right)^{-n/2}, \tag{34}$$

em que SQ_{Res} é a soma de quadrados de resíduos do modelo.

- Para o modelo restrito a verossimilhança maximizada fica dada por:

$$L_0 = \left(2\pi \frac{SQ_{Res_0}}{n}\right)^{-n/2}, \tag{35}$$

em que SQ_{Res_0} é a soma de quadrados de resíduos para o modelo restrito (ajustado apenas com as $k - q$ variáveis não restritas a zero.

Inferência para vários parâmetros do modelo

- O teste da razão de verossimilhanças para testar H_0 baseia-se na estatística da razão de verossimilhanças:

$$\frac{L_0}{L} = \left(\frac{SQ_{Res_0}}{SQ_{Res}}\right)^{-n/2} = \left(\frac{SQ_{Res}}{SQ_{Res_0}}\right)^{n/2}. \tag{36}$$

- Sob a hipótese H_0, assintoticamente:

$$\Lambda = -2\ln\left(\frac{L_0}{L}\right) \sim \chi^2_q, \tag{37}$$

em que χ^2_q denota a distribuição qui-quadrado com q graus de liberdade.

- Para um nível de significância α, H_0 será rejeitada se Λ superar o quantil $1-\alpha$ da distribuição χ^2_q.

Inferência para vários parâmetros do modelo

- No caso de modelos lineares, no entanto, temos um teste exato como alternativa ao teste assintótico baseado na distribuição qui-quadrado;
- Sob a hipótese nula, a estatística:

$$F_0 = \frac{(SQ_{Res_0} - SQ_{Res})/q}{SQ_{Res}/(n-p)} \tag{38}$$

tem distribuição F-Snedecor com q e $n-p$ graus de liberdade.

- Observe que F_0 baseia-se na variação da soma de quadrados de resíduos resultante da restrição aplicada a q parâmetros do modelo.
- A hipótese H_0 deverá ser rejeitada, ao nível de significância α, se F_0 superar o quantil $1-\alpha$ da distribuição F-Snedecor com q e $n-p$ graus de liberdade.

Inferência para vários parâmetros do modelo

- A estatística F_0 pode ser calculada por:

$$F_0 = \frac{(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^{(0)})' \boldsymbol{V}_{11}^{-1} (\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^{(0)})}{qQM_{Res}}, \tag{39}$$

em que $\hat{\boldsymbol{\beta}}_q$ denota o vetor de q entradas de $\hat{\boldsymbol{\beta}}$ referente aos parâmetros restritos e $\boldsymbol{V}_{11}$ a matriz quadrada com as q entradas (linhas e colunas) de $\boldsymbol{(X'X)}^{-1}$.

- Repare que nesta representação $\boldsymbol{\beta}_q^{(0)}$ representa o vetor postulado para os q parâmetros restritos sob H_0 (geralmente um vetor de zeros).

Inferência para vários parâmetros do modelo

- O teste da significância do modelo de regressão ($H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$) é um caso particular, em que a estatística F, apresentada no quadro da análise de variância, tem distribuição F-Snedecor com $p - 1$ e $n - p$ graus de liberdade sob H_0.

Região de confiança

- Suponha interesse em estimar simultaneamente algum subconjunto de parâmetros do modelo.
- Seja $\boldsymbol{\beta}_q$ um subconjunto de elementos de $\boldsymbol{\beta}$, contendo parâmetros sobre os quais se deseja inferir.
- Adicionalmente, seja $\hat{\boldsymbol{\beta}}_{\boldsymbol{q}}$ o vetor de estimadores de mínimos quadrados de $\boldsymbol{\beta}_q$.
- Uma região de confiança $100(1-\alpha)\%$ para os componentes de $\boldsymbol{\beta}_q$ é definido pelo conjunto de todos os vetores $\boldsymbol{\beta}_q^{(0)}$ tais que:

$$F_0 = \frac{(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)})' \boldsymbol{V}_{11}^{-1} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)})}{qQM_{Res}} \leq F_{q,n-p}(1-\alpha) \qquad (40)$$

em que $F_{q,n-p}(\alpha)$ é o quantil $1-\alpha$ da distribuição F-Snedecor com q e $n-p$ graus de liberdade.

Testes de hipóteses para combinações lineares dos parâmetros

- De forma mais geral, podemos definir hipóteses lineares na forma:

$$H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}, \tag{41}$$

em que $\boldsymbol{L}$ é uma matriz de constantes de dimensão $q \times p$, de rank linha completo, e $\boldsymbol{c}$ um vetor de constantes de dimensão q (ambos especificados).

- Neste caso, H_0 compreende q hipóteses lineares sobre os parâmetros do modelo, do tipo:

$$\begin{gathered}
L_{11}\beta_0 + L_{12}\beta_1 + L_{13}\beta_2 + ... + L_{1p}\beta_k = c_1 \\
L_{21}\beta_0 + L_{22}\beta_1 + L_{23}\beta_2 + ... + L_{2p}\beta_k = c_2 \\
\vdots \\
L_{q1}\beta_0 + L_{q2}\beta_1 + L_{q3}\beta_2 + ... + L_{qp}\beta_k = c_q
\end{gathered}$$

Testes de hipóteses para combinações lineares dos parâmetros

- Sob a hipótese H_0 a estatística:

$$F = \frac{(\boldsymbol{L}\beta - \boldsymbol{c})'[\boldsymbol{L}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{L}']^{-1}(\boldsymbol{L}\beta - \boldsymbol{c})}{qQM_{Res}} \tag{42}$$

tem distribuição F-Snedecor com q e $n-p$ graus de liberdade.

- Assim, a hipótese nula será rejeitada, ao nível de significância α, se o valor calculado da estatística F exceder o quantil $1-\alpha$ da distribuição F-Snedecor com q e $n-p$ graus de liberdade.

Intervalos de confiança para combinações lineares dos parâmetros

- Seja $\boldsymbol{l}' = (l_0, l_1, l_2, ..., l_p)$ um vetor de constantes e considere interesse em estimar $\theta = \boldsymbol{l}'\boldsymbol{\beta}$.
- A estimativa pontual de $\boldsymbol{l}'\boldsymbol{\beta}$ é dada por $\boldsymbol{l}'\hat{\boldsymbol{\beta}}$.
- Um intervalo de confiança $100(1-\alpha)\%$ para $\boldsymbol{l}'\boldsymbol{\beta}$ tem limites:

$$\boldsymbol{l}'\hat{\boldsymbol{\beta}} \mp t_{n-p,\alpha/2}\sqrt{\hat{\sigma}^2\boldsymbol{l}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{l}}. \quad (43)$$

Regressão com coeficientes padronizados

- Na análise de RLM, a comparação das magnitudes dos $\hat{\beta}_j'$s em sempre é possível devido ao impacto das diferentes unidades de medidas dos x_j's.
- Caso seja desejado que tais estimativas sejam comparáveis, pode-se padronizar cada uma das variáveis de forma que as variáveis resultantes tenham uma mesma escala.
- Uma allternativa de padronização consiste em 'normalizar' cada uma das variáveis, aplicando:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, 2, ..., n; \ j = 1, 2, ..., k, \tag{44}$$

e

$$y_i^* = \frac{y_i - \bar{y}}{s_y}, \quad i = 1, 2, ..., n, \tag{45}$$

Regressão com coeficientes padronizados

- Neste caso, $\bar{x}_j$ e s_j são a média e o desvio padrão amostrais de x_j e $\bar{y}$ e s_y a média e desvio padrão amostrais de y.
- Usando as variáveis normalizadas, o modelo de regressão linear múltipla fica definido por:

$$y_i^* = b_1 z_{i1} + b_2 z_{i2} + ... + b_k z_{ik} + \epsilon_i, \quad i = 1, 2, ..., n. \tag{46}$$

- A análise segue então da maneira usual de forma que o estimador de mínimos quadrados de $\boldsymbol{b}$ fica dado por:

$$\hat{\boldsymbol{b}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}. \tag{47}$$

Regressão com coeficientes padronizados

- Ao centrar as variáveis, o intercepto do modelo é deslocado para zero.
- As interpretações dos parâmetros do modelo devem ser feitas em termos dos valores padronizados das variáveis originais.