Modelo de regressão linear geral

Em geral, a variável resposta Y pode estar relacionada com p variáveis explicativas. O modelo de regressão é definido por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \qquad (2)$$

é conhecido como modelo de primeira ordem, pois é linear nos parâmetros e linear nas variáveis explicativas. Esse modelo é conhecido como um modelo de regressão linear múltipla ou também como um modelo de regressão linear geral.

- Y_i é o valor da variável resposta para a i-ésima observação,
- β_0, $\beta_1, \ldots, \beta_p$ são parâmetros desconhecidos,
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ são constantes conhecidas,
- ε_i são independentes e $N(0, \sigma^2)$, $i = 1, \ldots, n$.

Modelo de Regressão Linear Geral em termos matriciais

- Dessa forma, o modelo (2) pode ser escrito em termos matriciais:

$$\mathbf{Y}_{nx1} = \mathbf{X}_{nx(p+1)}\boldsymbol{\beta}_{(p+1)x1} + \boldsymbol{\varepsilon}_{nx1}$$

- As condições do erro do modelo (2) são:

$$E(\boldsymbol{\varepsilon})_{nx1} = \mathbf{0}_{nx1} \quad \text{e} \quad Var(\boldsymbol{\varepsilon})_{nxn} = \sigma^2\mathbf{I}_{nxn}$$

- Consequentemente,

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$$
$$Var(\mathbf{Y}) = Var(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{0} + Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

- O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é dada por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (3)$$

Estimador de σ^2

- Como definido anteriormente, o estimador de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)} = \frac{SQRes}{n-(p+1)} = MSRes$$

- Em termos matriciais:

$$\hat{\sigma}^2 = \frac{\mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}}{n-(p+1)} = \frac{SQRes}{n-(p+1)} = MSRes$$

Descrição dos dados

Em um estudo de inovação no setor de seguros, um economista desejava relacionar a velocidade com que uma determinada inovação de seguros é adotada (Y) para o tamanho da empresa de seguros (X_1) e o tipo de empresa (X_2).

Variáveis:

- Y - medida pelo número de meses decorridos,
- X_1 - medida em milhões de dólares,
- X_2 - tipo da empresa: empresa de ações e empresa comum.

Para escrever a variável qualitativa no modelo, poderia considerar duas variáveis indicadoras: X_2 e X_3, da seguinte forma:

$$X_2 = \begin{cases} 1 & \text{se empresa de ações} \\ 0 & \text{caso contrário} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{se empresa comum} \\ 0 & \text{caso contrário} \end{cases}$$

1) Escrever o modelo de regressão para representar os dados do estudo.

2) Suponha que você tem apenas 4 observações (n=4) no estudo e que para as duas primeiras o tipo de empresa é de ações, escreva a matriz $\mathbf{X}$.

3) Encontre $\mathbf{X}'\mathbf{X}$.

4) Analise a matriz $\mathbf{X}'\mathbf{X}$.

1) O modelo de regressão para os dados em estudo é descrito por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad (4)$$

2) $$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix}$$

3) $$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ X_{11} & X_{21} & X_{31} & X_{41} \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & \sum_{i=1}^{4} X_{i1} & 2 & 2 \\ \sum_{i=1}^{4} X_{i1} & \sum_{i=1}^{4} X_{i1}^2 & \sum_{i=1}^{2} X_{i1} & \sum_{i=3}^{4} X_{i1} \\ 2 & \sum_{i=1}^{2} X_{i1} & 2 & 0 \\ 2 & \sum_{i=3}^{4} X_{i1} & 0 & 2 \end{bmatrix}$$

Observação:

- Pode-se perceber que a primeira coluna da matriz $\mathbf{X}'\mathbf{X}$ é igual a soma das duas últimas colunas. Então, as colunas são linearmente dependentes.

- Dessa forma, a matriz $\mathbf{X}'\mathbf{X}$ não tem inversa. Logo, não tem estimadores únicos para os coeficientes de regressão.

Solução:

- Uma solução simples é retirar uma variável indicadora. Com essa solução, a interpretação dos coeficientes é simples.

$\Longrightarrow$ De maneira geral, uma variável qualitativa com c categorias, deve ser representada por $c - 1$ variáveis explicativas. Cada variável indicadora assumirá valores 0 ou 1.

- Ao retirar uma variável indicadora do modelo (4) que representa os dados em estudo, o modelo será descrito por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (5)$$

em que:

- X_i1 = tamanho da empresa (em milhões de dólares)
- $X_{i2} = \begin{cases} 1 & \text{se empresa de ações} \\ 0 & \text{se empresa comum} \end{cases}$

- A função resposta para o modelo (5) é:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Para entender o significado dos coeficientes no modelo, considere a situação que $X_2 = 0$, a função resposta é dada por:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2(0) = \beta_0 + \beta_1 X_1 \quad \text{empresa comum}$$

Assim, a função resposta para a empresa comum é uma reta com intercepto β_0 e coeficiente angular β_1.

- Para $X_2 = 1$, a função resposta é:

  $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_1$ empresa de ações

  Assim, a função resposta para a empresa de ações também é uma reta com intercepto $\beta_0 + \beta_2$ e coeficiente angular β_1.

Interpretação: β_2 indica que a função resposta, ou a resposta média, para as empresas de ações é maior (ou menor) do que a resposta média para as empresas comuns, para qualquer tamanho de empresa.

Interpretação geral: β_2 mostra quanto maior (ou menor) é a resposta média para a categoria codificada como 1 do que a resposta média para a categoria codificada como 0, para qualquer valor de X_1.

Analisar os dados no R

Critérios para seleção de modelos

- Para um conjunto de dados com p variáveis explicativas, pode-se construir 2^p modelos.
- Esse cálculo considera que cada variável explicativa pode ser incluída ou excluída do modelo.
- Assumi-se que o número de observações exceda o número de parâmetros. Logo,

$$n > p + 1$$

Critérios para seleção de modelos

- Alguns critérios são: R^2_{p+1}, $R^2_{a,p+1}$, C_{p+1}, AIC_{p+1}, BIC_{p+1} e $PRESS_{p+1}$.

- Métodos automáticos.

Critério R^2_{p+1} ou $SQRes_{p+1}$

- Deve-se examinar o Coeficiente de determinação multiplo dos modelos para encontrar o ponto em que adicionar mais variáveis não é útil, pois o aumento de R^2_{p+1} é pequeno.

$$R^2_{p+1} = 1 - \frac{SQRes_{p+1}}{SQT}$$

- Como a SQT é a mesma para todos os modelos de regressão possívies, examinar o aumento de R^2_{p+1} é equivalente a examinar a diminuição em $SQRes_{p+1}$.
- Para identificar o ponto em que adicionar mais variáveis não é útil, pode-se fazer um gráfico do número de parâmetros presente no modelo versus o respectivo valor de R^2_{p+1}.

Critério $R^2_{a,p+1}$ ou $MSRes_p$

- Como no calculo do R^2_{p+1} não é considerado o número de parâmetros do modelo e desde que $\max(R^2_{p+1})$ não diminui com o aumento do número de parâmetros no modelo, uma sugestão é utilizar o coeficiente de determinação múltiplo ajustado:

$$R^2_{a,p+1} = 1 - \frac{\frac{SQRes_{p+1}}{n-(p+1)}}{\frac{SQT}{n-1}} = \frac{MSRes_{p+1}}{MST}$$

- Para identificar o ponto em que adicionar mais variáveis não é útil, pode-se fazer um gráfico do número de parâmetros presente no modelo versus o respectivo valor de $R^2_{a,p+1}$.

Critério C_{p+1} de Mallows's

- O Critério C_{p+1} de Mallows's é definido por:

$$C_{p+1} = \frac{SQRes_{p+1}}{MSRes} - (n - 2p - 2)$$

em que

- $SQRes_{p+1}$ é a soma de quadrados do resíduo do modelo de regressão ajustado com as p variáveis explicativas, ou seja, $p + 1$ parâmetros.
- $MSRes$ é o quadrado médio do resíduo para modelo com todas as variaveis.

Critério C_{p+1} de Mallows's

- Se o modelo com p variáveis é adequado, isto é, o uso de apenas p variáveis explicativas equivale ao uso de todas, temos que

$$MSRes_{p+1} = \frac{SQRes_{p+1}}{n-(p+1)}$$

é um estimador de σ^2, assim como $\hat{\sigma}^2$. Sendo assim, pode-se provar que:

$$E(C_{p+1}) \cong p+1$$

- Os critérios $R^2_{a,p+1}$ e C_{p+1} penalizam modelos com um número grande de variáveis explicativas.

- Alternativas para esses critérios, mas que também fornecem penalidades para a adição de variáveis explicativas são os critérios de Akaike e Critério de Informação Bayesiano.

Critério de Akaike - AIC_{p+1}

- Esse critério foi proposto por Akaike (1974) que utilizou a informação de Kullback-Leiber para definir o critério AIC_{p+1}.
- A informação de Kullback-Leiber (K-L) é uma medida da distância entre o modelo verdadeiro e um modelo candidato.
- Akaike desenvolveu uma estimativa da medida de K-L, baseada no logarítmo da função de verossimilhança no ponto de máximo, acrescida de uma penalidade associada ao número de parâmetros.
- A função de penalidade tem a finalidade de corrigir um viés proveniente da comparação de modelos com diferentes números de parâmetros

Critério de Akaike - AIC_{p+1}

- A medida denominada de informação de Akaike é definida por:

$$AIC_{p+1} = -2ln(L(\hat{\theta})) + 2(p+1),$$

em que $p+1$ é o número de parâmetros estimados no modelo.

- Usando regressão de mínimos quadrados ordinários, tem-se que:

$$AIC_{p+1} = n * ln(SQRes_{p+1}) - n * ln(n) + 2(p+1),$$

- O modelo escolhido deve apresentar menor valor de AIC_{p+1} dentre todos os modelos considerados para determinado problema.

- Burnham e Anderson (2002) recomendam o uso do AIC_{p+1} apenas quando $n/(p+1) \geq 40$

Critério de Akaike corrigido - $AICc_{p+1}$

- Para pequenas amostras, $n/(p+1) < 40$
- Hurvich e Tsai (1989) desenvolveram o seguinte critério:

$$AICc_{p+1} = AIC + \frac{2(p+1)(p+2)}{n-(p+1)-1} = AIC + \frac{2(p+1)(p+2)}{n-p-2},$$

- O modelo escolhido deve apresentar menor valor de AICc dentre todos os modelos considerados para determinado problema.

Critério de informação Bayesiano - BIC_{p+1}

- O Critério de informação Bayesiano(BIC), proposto por Schwarz (1978), é definido por:

$$BIC_{p+1} = -2log(L(\hat{\theta})) + (p+1) * ln(n)$$

- Usando regressão de mínimos quadrados ordinários, tem-se que:

$$BIC_{p+1} = n * ln(SQRes_{p+1}) - n * ln(n) + (p+1) * ln(n)$$

- O modelo escolhido deve apresentar menor valor de BIC_{p+1} dentre todos os modelos considerados para determinado problema.

- O critério BIC_{p+1} penaliza mais modelos com maior número de parâmetros do que o critério AIC_{p+1}. Tendendo dessa forma, a selecionar modelos com um número menor de parâmetros.

Métodos automáticos

Eles podem ser classificados como:

1) Seleção "Forward"

2) Eliminação "Backward"

3) Regressão "Stepwise"

Seleção "Forward"

Passo1 Ajusta modelos de regressão simples com cada variável explicativa. Por meio da estatística F e seu respectivo p-valor, seleciona a variável mais significativa.

Passo2 Considera a variável selecionada no Passo 1 e ajusta todos os possíveis modelos de regressão com duas variáveis, sendo a variável selecionada no Passo 1 presente em todos os modelos. Para cada modelo de regressão calcula a estatística F parcial para verificar o efeito de introduzir a variável X_k no modelo que já tem a variável selecionada no Passo1. Se alguma variável for selecionada o processo continua. Caso contrário, o processo termina.

Seleção "Forward"

Passo3 Supõe que alguma variável foi selecionada no Passo2. Em seguida ajusta-se todos os possíveis modelos com três variáveis, sendo as variáveis selecionadas nos Passos 1 e 2 presentes nos modelos com três variáveis. Para cada modelo de regressão calcula a estatística F parcial para verificar o efeito de introduzir a variável X_k no modelo. Se alguma variável for selecionada o processo continua. Caso contrário, o processo termina.

Passo4 Repete o procedimento até o algoritmo não selecionar mais variáveis que devem entrar no modelo.

Eliminação "Backward"

Esse procedimento é o oposto da Seleção "Forward".

Passo1 Ajusta-se um modelo com todas as variáveis explicativas e calcula a estatística F parcial e respectivo p-valor para cada variável do modelo. Se algum p-valor **for maior** que α estabelecido, então a variável é removida do modelo. Caso contrário, termina o procedimento.

Passo2 Se foi removida alguma variável no Passo1, agora ajusta-se um modelo com todas as variáveis que ficaram no modelo e calcula a estatística F parcial e respectivo p-valor para cada variável do modelo. Se algum p-valor **for maior** que α estabelecido, então a variável é removida do modelo. Caso contrário, termina o procedimento.

Eliminação "Backward"

Passo3 Repete o procedimento até o algoritmo não identificar mais variáveis que devem ser retiradas do modelo.

Regressão "Stepwise"

- Esse procedimento é semelhante ao Seleção "Forward", porém tem um passo que verifica se as variáveis que já estavam no modelo devem ou não continuar.

Regressão "Stepwise"

Passo1 Ajusta modelos de regressão simples com cada variável explicativa. Por meio da estatística F e seu respectivo p-valor, seleciona a variável mais significativa.

Passo2 Considera a variável selecionada no Passo 1 e ajusta todos os possíveis modelos de regressão com duas variáveis, sendo a variável selecionada no Passo 1 presente em todos os modelos. Para cada modelo de regressão calcula a estatística F parcial para verificar o efeito de introduzir a variável X_k no modelo que já tem a variável selecionada no Passo1. Se alguma variável for selecionada o processo continua. Caso contrário, o programa termina.

Regressão "Stepwise"

Passo3 Supõe que alguma variável foi selecionada no Passo2. Nesse passo será verificado se a variável selecionada no Passo1 deve permanecer no modelo formado pelas variáveis selecionadas nos Passos 1 e 2. E por meio da estatística F verifica se a variável selecionada no Passo 1 fica ou sai do modelo.

Passo4 Supõe que as variáveis selecionadas nos Passos 1 e 2 ficam no modelo. O próximo passo é verificar qual é a próxima variável que deverá entrar no modelo e em seguida verifica-se qual(is) variável(is) que já estavam no modelo ficará ou sairá. Esses passos vão se repetindo até chegar no modelo final.

Implementar os critérios definidos para os dados em estudo