Planejamentos Experimentais em Modelos de Regressão

Bárbara Mariotti (13686473) e Guilherme Freitas (13686386)

Introdução

- A análise de regressão é uma técnica estatística usada para pesquisar e modelar a associação de diversas variáveis de um processo.

- O planejamento de experimentos tem algumas estratégias como:

    - Reconhecer o problema

    - Identificar possíveis fatores que podem afetar esse problema

    - Verificar quais fatores poderão ser mantidos fixos

    - Escolher um projeto experimental adequado

    - Escolher a resposta adequada

Objetivos do Planejamento Experimental

- Planejamento tem como função auxiliar na escolha dos valores regressores.

- Consideremos o modelo linear geral $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ onde:

    - $\mathbf{Y} = [\, Y_1, \dots , Y_n \,]^T$ um vetor aleatório de observações

    - $\mathbf{X} = [\, X_{ij} \,]$ uma matriz de dimensão n $\times$ (p+1) contendo os valores das p variáveis regressoras e uma coluna de 1_s

    - $\boldsymbol{\beta} = [\, \beta_0, \dots , \beta_p \,]^T$ um vetor de parâmetros desconhecidos

    - $\boldsymbol{\varepsilon} = [\, \varepsilon_1, \dots , \varepsilon_n \,]^T$ um vetor aleatório de erros

Objetivos do Planejamento Experimental

- Critérios desejáveis para um planejamento experimental:
    - Que todos os pontos exerçam igual influência na determinação das estimativas dos coeficientes de regressão.
    - Minimização de algumas funções das variâncias dos estimadores
    - O planejamento deveria permitir detectarmos a necessidade de inclusão de termos não lineares

Critérios de planejamento

- Podemos ajustar modelos levando em conta vários critérios.

- Um desses critérios é que o planejamento seja pouco sensível a observações discrepantes.

- Uma possível medida para essa sensibilidade é a soma de quadrados da diagonal principal da matriz $X(X^T X)^{-1}X^T$

Demais propriedades

I) Gerar uma distribuição satisfatória de informação através de uma região de interesse R.

II) Assegurar que o valor ajustado em x, $\hat{y}(x)$ seja tão próximo quanto possível do valor real . em x, E(y|x).

III) Fornecer a possibilidade de detectar falta de ajuste.

IV) Permitir que transformações sejam estimadas.

V) Permitir que os experimentos sejam realizados em blocos quando necessário.

VI) Permitir que planejamentos de ordem crescente sejam construídos sequencialmente.

VII) Proporcionar uma estimativa interna da variância do erro.

Demais propriedades

VIII) Ser insensível ou pouco sensível a observações discrepantes e a violação de suposições da teoria normal usual.

IX) Requerer um número mínimo de pontos experimentais.

X) Proporcionar padrões simples de distribuição dos dados que permitam apreciações visuais rápidas.

XI) Assegurar simplicidade de cálculo.

XII) Comportar-se bem na presença de erros no conjunto das variáveis preditoras.

XIII) Não requerer um número muito grande de níveis das variáveis preditoras.

XIV) Proporcionar posteriormente uma avaliação da suposição de igualdade de variâncias (homocedasticidade).

Critérios de planejamento

É desejável que o estatístico aplicado tenha habilidade de perceber as necessidades especiais de cada situação e escolher o melhor planejamento.

Critérios de planejamento

Supõe-se que um pesquisador quer coletar dados de uma variável resposta Y associados à n valores selecionados de uma variável preditora controlável para determinar uma relação entre a variável Y e a variável preditora.

Nessas condições cabe avaliar:

Critérios de planejamento

1) Em qual intervalo de valores da variável preditora o pesquisador está realmente interessado?

   O intervalo deve ser amplo o suficiente para permitir inferência útil e uma vez que a decisão é tomada, o intervalo pode ser codificado para se tornar [-1,1], sem perda de generalidade. Em geral a transformação que permite tal codificação é dada por:

$$X = \frac{\text{variável original} - \text{ponto médio do intervalo original}}{\text{metade da amplitude do intervalo original}}$$

Critérios de planejamento

2) Que tipo de de relação o pesquisador espera obter no intervalo selecionado? Linear? Quadrática? Ou outra?

   O pesquisador pode usar o próprio conhecimento ou consultar especialistas.

3) Verificou-se que o modelo de 1º grau é inadequado, então qual modelo o pesquisador espera como alternativa?

   Se sua crença inicial é que o modelo verdadeiro é uma reta, se isso não se mostrar adequado, possivelmente espera-se um modelo quadrático, ou mais remotamente, cúbico.

Critérios de planejamento

4) O pesquisador deverá incorporar réplicas nos experimentos de modo que σ^2 possa ser estimado através do erro puro e, dessa forma, as suposições usuais, principalmente a de homocedasticidade, possam ser checadas.

5) Quantos dados experimentais são possíveis?

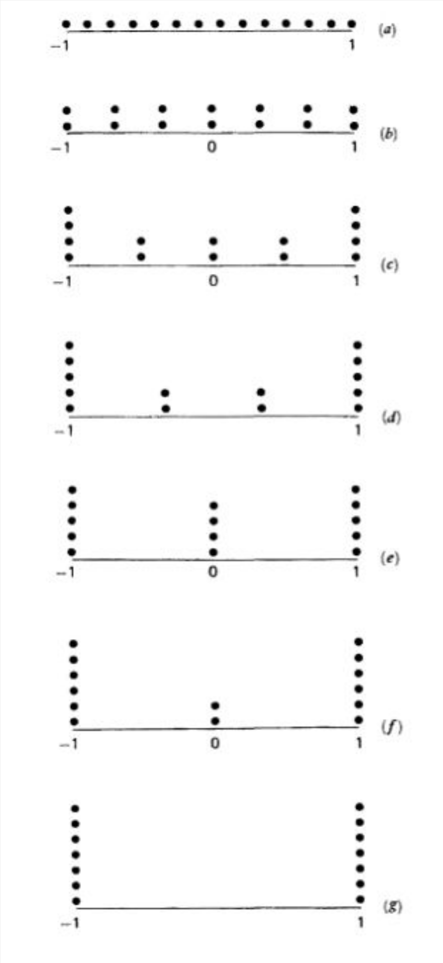
   Geralmente o pesquisador tem problemas de tempo, dinheiro, ou até mesmo motivos práticos que dificultam a obtenção dos dados.

6) Quantas posições, ou seja, diferentes valores de X, deveriam ser escolhidos e quantas observações devem ser obtidas para cada valor de X?

Exemplo ilustrativo

Supondo que o pesquisador determinou que a relação de 1º grau é correta no intervalo $-1 \leq X \leq 1$ do seu preditor codificado. Não conhecemos σ^2 e temos várias opções de, por exemplo, posições para uma amostra de $n = 14$:

Exemplo ilustrativo



Cada uma das possibilidades tem 14 graus de liberdade, sendo 2 deles para a estimação de β_0 e β_1, restando 12 graus de liberdade para serem distribuídos entre falta de ajustamento e erro puro.

Exemplo ilustrativo

Ajustando o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, na presença de réplicas, as somas de quadrados da falta de ajuste e do erro puro são dadas por:

$$SQFA = \sum_{i=1}^{k} n_i (\bar{y}_i - \hat{y}_i)^2 \text{ e } SQEP = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Sendo que

k: número de posições

$\bar{y}_i$: média dos valores da variável Y associados ao i-ésimo valor x_i da variável X

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, estimador da média $E(y \mid x_i) = \beta_0 + \beta_1 x_i$

$\hat{\beta}_0$ e $\hat{\beta}_1$ estimadores de mínimos quadrados de β_0 e β_1

n_i : número de observações associadas ao valor x_i da variável X, i=1,2,...,k .

Exemplo ilustrativo

Além disso, a soma de quadrados do resíduo, SQRes, é definida como:

$$SQ\mathrm{Re}s = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2$$

E verifica-se que SQRes = SQEP + SQFA.

Exemplo ilustrativo

Fonte de Variação	GL	SQ	QM	F
Falta de ajuste	k-2	SQFA	QMFA	$F_{FA} = \dfrac{QMFA}{QMEP}$
Erro Puro	n-k	SQEP	QMEP	
Resíduo	n-2	SQRes		

A tabela corresponde à análise de variância associada ao teste de falta de ajustamento, que testa H_0: O modelo é linear; e rejeita a hipótese se:

$$F_{FA} = \frac{\dfrac{SQFA}{k-2}}{\dfrac{SQEP}{n-k}} \geq F_{c,1-\alpha}$$

Em que $F_{c,1-\alpha}$ é o quantil de ordem 1-α da distribuição F com k-2, n-k graus de liberdade.

Exemplo ilustrativo

	(a)	(b)	(c)	(d)	(e)	(f)	(g)
(1)graus de liberdade de falta de ajuste	12	5	3	2	1	1	0
(2)graus de liberdade do erro puro	0	7	9	10	11	11	12
(3) $DP(\hat{\beta}_1)/\sigma$ 0,27	0,43	0,40	0,33	0,31	0,32	0,29	
(4) k	14	7	5	4	3	3	2

Já essa tabela apresenta algumas características associadas às sete estratégias de separação da amostra, e que são fundamentais para a escolha de uma delas.

Exemplo ilustrativo

- (a) é uma estratégia pobre, se desejarmos testar a falta de ajuste, (g) também é eliminada.

- (b) tem níveis demais se nossa alternativa é um modelo quadrático. Além de ter o maior valor de $DP(\beta) / \sigma$, dentre os planejamentos restantes.

- A rigor, somente três níveis são estritamente necessários num teste de falta de ajuste do modelo de grau um contra uma alternativa quadrática, mas, neste caso existiria somente um grau de liberdade para falta de ajuste em (e) e (f).

- O planejamento (d) possui dois graus de liberdade para falta de ajuste, já o planejamento (c) talvez tenha um número excessivo de níveis.

- Então a escolha final se encontra entre (f) e (d), com (f) sendo levemente preferida se a alternativa quadrática é a única esperada.

Exemplo ilustrativo

Entretanto, o aspecto mais importante dessa discussão não é a escolha específica de um planejamento, mas sim a eliminação imediata de planejamentos que poderiam, em outros contextos, ser considerados razoáveis.

Referências

- https://www.teses.usp.br/teses/disponiveis/45/45133/tde-02062009-185614/publico/CamilaSerinoli.pdf

- https://drive.google.com/file/d/118MHS6jS61hpdq-HIZYFJXZDAQqmH48x/view