

3. Análise de Variância

3.1 O modelo com uma amostra

Seja Y_1, \dots, Y_n iid's com

$$Y_i \sim N(\beta, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

\Rightarrow média e variância comuns para as n observações.

\Rightarrow testes de hipóteses sobre β e σ^2 .

Se $\mu = (\mu_1, \dots, \mu_n)^T$ denota o vetor de média para $Y = (Y_1, \dots, Y_n)^T$ então, sob o modelo (1) μ tem a forma $\beta \underline{1}$ onde $\underline{1} = (1, \dots, 1)^T$. Assim, o modelo (1) é um modelo linear dado pelo subespaço linear $L_1 = \text{span}\{\underline{1}\}$.

Para estimar os parâmetros deste modelo, calculamos a projeção ortogonal de y sobre L_1 dada por

$$p_1(Y) = \frac{1 \cdot Y}{\|1\|^2} 1 = \bar{Y}_+ 1.$$

Em particular, o estimador para β é $\hat{\beta} = \bar{Y}_+$ e, a distribuição de $\hat{\beta}$ sob L_1 é $N(\beta, \sigma^2/n)$.

O estimador para σ^2 sob L_1 é

$$\tilde{\sigma}_1^2 = \frac{1}{n-1} \|Y - p_1(Y)\|^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_+)^2,$$

e a distribuição de $\tilde{\sigma}_1^2$ sob L_1 é dada por

$\tilde{\sigma}_1^2 \sim \sigma^2 \frac{\chi_{(n-1)}^2}{(n-1)}$. Além disso, $\tilde{\sigma}_1^2$ e $\hat{\beta}$ são independentes.

Pelo Teorema de Pitágoras, obtemos

$$\|y - p_1(y)\|^2 = \|y\|^2 - \|p_1(y)\|^2 = S_y - \hat{\beta}^2 \|1\|^2 = S_y - n\bar{y}_+^2$$

Teste t para a hipótese $H_2 : \beta = \beta_0 \in \mathbb{R}$.

Vimos que a estatística para o teste t é

$$t(Y) = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})},$$

onde $s.e.(\hat{\beta}) = \tilde{\sigma}_1 / \|\underline{1}\| = \tilde{\sigma}_1 / \sqrt{n}$.

Assim, $t(Y) = \sqrt{n}(\bar{Y}_+ - \beta_0) / \tilde{\sigma}_1$ e $t(Y) \sim t_{(n-1)}$ sob H_2 .

Um intervalo de $1 - \alpha$ de confiança para β é dado por

$$IC(\beta, 1 - \alpha) : \bar{Y}_+ \pm t_{1-\frac{\alpha}{2}(n-1)} \frac{\tilde{\sigma}_1}{\sqrt{n}}.$$

Similarmente ao modelo de regressão linear simples, as condições para o modelo (1) são:

(i) as médias dos Y_i 's são constantes, (ii) as variâncias dos Y_i 's são constantes, (iii) a distribuição de Y_i é normal, (iv) as variáveis Y_1, \dots, Y_n são independentes.

As condições (i), (ii) e (iv) são mais difíceis de serem verificadas a partir dos dados. Em geral estas condições são consequências das especificações do experimento.

A condição (iv) requer n experimentos individuais separados no tempo e no espaço e, (i) e (ii) exigem que as condições do experimento permaneçam constantes para os n experimentos.

A condição (iii) pode ser verificada por um “normal-plot” das n observações. Porém, como a linearidade do normal-plot também requer (i), (ii) e (iv), o mesmo é , implicitamente, uma verificação das quatro condições.

Os resíduos $r_i = y_i - \hat{\beta}$ são uma transformação linear das observações y_i 's tal que um normal-plot dos mesmos é equivalente a um normal plot dos y_i 's.

EXEMPLO 1: Pesos em (onças) para 15 pacotes de açúcar.

16.1	15.8	15.8	15.9	16.1
16.2	16.0	15.9	16.0	15.7
15.7	15.8	16.0	16.0	15.8

O fabricante garante que cada pacote contém em média 16 onças de açúcar.

Parece então razoável aplicar o modelo (1) para estes dados.

Supomos então Y_1, \dots, Y_{15} iid's $N(\beta, \sigma^2)$.

Assim temos, $\hat{\beta} = \bar{y}_+ = 15.92$ e $\tilde{\sigma}_1^2 = 0.02314$.

No teste de $H_2 : \beta = 16$, temos $t(y) = -2.037$ o que dá um P-valor de 0.06 (teste bilateral). Assim, ao nível de significância de 5% não podemos rejeitar H_2 . Um resultado mais forte exige uma amostra maior.

$$IC(\beta, 0.95) : (15.836, 16.004)$$

Se H_A fosse $\beta < 16$, o P-valor seria 0.03, sugerindo que a estatística de teste é significativa. Por outro lado, a hipótese $H_A : \beta > 16$ estaria protegendo o fabricante sistematicamente, levando a um P-valor neste caso de 0.97. Neste caso, como ambas as H_A 's unilaterais parecem importantes, parece preferível usar o teste bilateral.

3.2 Observações emparelhadas

⇒ Generalização do caso anterior para amostras emparelhadas.

Suponha aqui $Y_i = U_{i2} - U_{i1}$, $i = 1, \dots, n$.

Suponha também que os n pares (U_{i1}, U_{i2}) , são independentes com

$U_{i1} \sim N(\delta_i, \tau^2)$ e $U_{i2} \sim N(\delta_i + \beta, \tau^2)$, $i = 1, \dots, n$.

Finalmente, suponha que a correlação entre U_{i1} e U_{i2} é $\rho_{12} = \rho$, $i = 1, \dots, n$.

Se cada par (U_{i1}, U_{i2}) segue uma distribuição normal bivariada, segue que $Y_i \sim N(\beta, \sigma^2)$, onde $\sigma^2 = 2\tau^2(1 - \rho)$.

Assim, as diferenças Y_i , $i = 1, \dots, n$ são independentes e identicamente distribuídas.

Podemos então, testar $H_2 : \beta = 0$ que é a hipótese representando que não há diferença na média para as duas observações de cada par.

O modelo aqui então é exatamente o modelo estudado na seção anterior. Assim, os estimadores para β e σ^2 são

$$\hat{\beta} = \bar{y}_+ = \bar{u}_{2+} - \bar{u}_{1+} \text{ e } \tilde{\sigma}_1^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y}_+)^2.$$

Para verificar o modelo construímos os seguintes gráficos: um normal-plot das diferenças y_i e um diagrama de dispersão de u_{ij} versus i . No primeiro avaliamos a normalidade e, no segundo, se é razoável a suposição de variância comum.

A partir dos estimadores, podemos realizar o teste t para a hipótese $H_2 : \beta = 0$, dado por $t(y) = \frac{\sqrt{n}\bar{y}}{\hat{\sigma}_1}$ onde, sob H_2 , $t(Y) \sim t_{(n-1)}$.

(EXEMPLO 3.2)

Observação: O teste para observações emparelhadas pode ser usado mesmo que os U'_{ij} s não sejam normais. Basta que as diferenças Y_i sejam normais e independentemente distribuídas.

3.3 Análise de Variância com um fator

3.3.1 O Modelo

Seja $I \in \mathbb{R}^n$ um vetor tal que os valores de I pertençam ao conjunto $\{1, \dots, k\}$ para algum $k \in \mathbb{N}$. Para evitar trivialidades supomos que cada um dos valores $1, \dots, k$ ocorre pelo menos uma vez em I .

Tal vetor I é chamado um fator com k níveis. Para $k = 1$, temos o fator trivial $\underline{1} = (1, \dots, 1)^T$, que representa o caso onde os dados consistem de um único grupo.

Veremos agora o caso do modelo com um fator I onde a média seja constante dentro de cada grupo, mas pode ser diferente de grupo para grupo (modelo de análise de variância a um fator).

Será conveniente representarmos os dados por dois índices i denotando grupo e j denotando o número da observação dentro do grupo.

Assim, sejam Y_{ij} , $i = 1, \dots, k$, $j = 1, \dots, n_i$, $n_i \geq 1$ onde $\forall i, j$

$$Y_{ij} \sim N(\beta_i, \sigma^2)$$

O número total de observações é $n = \sum_i n_i$.

Observe que o índice j representa um fator J mas que este é um fator aninhado “nested” porque a observação j para um nível de I em geral não tem nada haver com a observação j para um outro nível de I . Em particular, nem todos os níveis de J precisam estar presentes para cada nível de I pois os n_i 's podem ser desiguais.

Este tipo de dado pode ser representado através do diagrama de dispersão de Y_{ij} versus i .

No modelo aqui proposto, cada grupo tem uma média diferente β_i , enquanto que a variância é comum. A questão principal a ser respondida pela análise estatística é se existe qualquer diferença entre as k médias β_1, \dots, β_k .

Este modelo está relacionado com o modelo de regressão linear simples no sentido de que ele descreve uma relação entre uma variável resposta Y e uma variável explicativa representada pelo fator I . Em alguns casos, os níveis de I representam os valores de algumas das variáveis subjacentes, por exemplo, uma discretização de uma variável contínua. Neste caso, o modelo representa uma relação completamente arbitrária entre esta variável e Y . Neste sentido, o modelo a um fator é mais flexível do que o modelo de regressão linear simples, que requer que a relação entre duas variáveis seja linear.

Para $\mu_{ij} = E(Y_{ij})$, temos, $\forall i, j$, $\mu_{ij} = \sum_{l=1}^k x_{ijl}\beta_l$, onde

$$x_{ijl} = \begin{cases} 1, & \text{se } l = i, \quad j = 1, \dots, n_i \\ 0, & \text{se } l \neq i, \quad j = 1, \dots, n_i \end{cases}$$

Isto mostra que o modelo é linear. O vetor de média μ é dado por

$$(\mu_{11}, \dots, \mu_{1n_1}, \dots, \mu_{k1}, \dots, \mu_{kn_k})^T \in \mathbb{R}^n.$$

O modelo é, então, representado pelo espaço linear $L = 1 = \text{span}\{e_1, \dots, e_k\}$ onde

$e_i = (0, \dots, 0, 1, \dots, 1, 0, \dots, 0)^T$, $i = 1, \dots, k$, onde os 1's aparecem no i -ésimo grupo. Os vetores e_1, \dots, e_k formam uma base para L_1 tal que $\dim L_1 = k$.

Observe que os vetores e_1, \dots, e_k são ortogonais e que $e_i \cdot e_i = n_i$. Estes vetores costumam ser chamados de vetores “dummy” associados ao fator I .

Devido à ortogonalidade, a projeção sobre L_1 é dada por

$$p_1(y) = \sum_{i=1}^k \frac{e_i \cdot y}{\|e_i\|^2} e_i = \sum_{i=1}^k \bar{y}_{i+} e_i$$

Assim, $\hat{\beta}_i = \bar{y}_{i+}$, $i = 1, \dots, k$.

A ortogonalidade dos vetores e_1, \dots, e_k implica na independência dos estimadores $\hat{\beta}_i$, $i = 1, \dots, k$, e, usando resultados anteriores, obtemos $\hat{\beta}_i \sim N(\beta_i, \frac{\sigma^2}{n_i})$, $i = 1, \dots, k$.

O estimador de σ^2 sob H_1 é

$$\tilde{\sigma}_1^2 = \frac{1}{n - k} \|y - p_1(y)\|^2$$

$$\tilde{\sigma}_1^2 \sim \sigma^2 \frac{\chi_{(n-k)}^2}{n-k}.$$

Pelo teorema de Pitágoras temos que $\|y - p_1(y)\|^2 = \|y\|^2 - \|p_1(y)\|^2 = S_y - \sum_i \frac{1}{n_i} y_{i+}^2$.

Temos então uma fórmula alternativa para $\tilde{\sigma}_1^2$ que corresponde a forma tradicional de calcular a variância amostral, envolvendo somente somas e somas de quadrados das observações.

3.3.2 Verificação do modelo

Usaremos aqui principalmente métodos gráficos para a verificação de modelos. Já usamos o normal-plot dos resíduos e, métodos adicionais serão apresentados adiante. Algumas vezes usaremos métodos numéricos, na forma de testes. Mas, métodos gráficos são frequentemente mais versáteis na prática.

Isto se deve ao fato de que muitos tipos de desvios podem ser percebidos a partir de um simples gráfico, mesmo padrões não previstos, enquanto que os testes podem ser “cegos” para outros desvios diferentes daqueles para o qual o teste foi designado.

O principal gráfico aqui é o diagrama de dispersão Y_{ij} versus i (padrões e variações na variância).

A questão da normalidade pode ser verificada de duas formas. Se os n_i 's são grandes tal que cada grupo contenha pelo menos de 10 a 15 observações, pode-se construir um normal-plot para cada grupo. Ao olharmos os gráficos, podemos buscar por padrões de desvios da normalidade que estejam presentes na maioria dos casos ou em todos os gráficos.

Um desvio que ocorra em apenas um gráfico normalmente não seria razão para rejeitarmos a hipótese de normalidade.

Alternativamente, pode-se fazer um único normal-plot para os resíduos definidos por

$$r_{ij} = y_{ij} - \bar{y}_{i+}$$

Um normal-plot dos n resíduos forneceria uma verificação global da normalidade e, como este gráfico é baseado em muitos valores, ele geralmente mostra um padrão mais claro do que os gráficos para grupos isolados.

Os resíduos podem ser úteis para outros tipos de gráficos. Um deles é gráfico de r_{ij} versus i . Os pontos neste gráfico devem ser distribuídos dentro de uma banda horizontal simétrica em torno do zero.

Desvio típico da homogeneidade das variâncias \rightarrow as variâncias crescem com os β_i 's. Um gráfico adequado para revelar tal padrão é obtido plotando-se r_{ij} versus $\hat{\beta}_i$.

modelo satisfatório \rightarrow pontos distribuídos em torno de uma banda horizontal. Quando detectarmos que o valor de σ^2 não é constante de grupo para grupo, devemos transformar Y_{ij} por $\log Y_{ij}$, por exemplo, e analisar as observações transformadas, verificando se agora a variância é de fato constante.

3.3.3 Teste de igualdade das médias

Uma questão fundamental na análise do modelo a um fator é teste para a hipótese

$H_2 : \beta_1 = \dots = \beta_k$. Sob H_2 , a média μ_{ij} é dada por β , onde β é o valor comum (desconhecido) de β_1, \dots, β_k .

Assim, a hipótese H_2 é um modelo linear, dado pelo subespaço linear $L_2 = \text{span}\{\underline{1}\}$, correspondendo ao modelo analisado na seção 3.1.

Escreveremos a seguir os resultados da seção 3.1 incorporando a notação com dois índices.

$$\hat{\beta} = \frac{1}{n} \sum_i \sum_j Y_{ij} = \bar{Y}_{++},$$

correspondendo à $p_2(Y) = \bar{Y}_{++}\underline{1}$ e

$$\tilde{\sigma}_1^2 = \frac{1}{n-1} \sum_i \sum_j (Y_{ij} - \bar{Y}_{++})^2.$$

Sob H_2 , $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{n})$ e $\tilde{\sigma}_1^2 \sim \sigma^2 \frac{\chi^2(n-1)}{n-1}$.

Para testar L_2 sob L_1 usamos o teste F definido por

$F(Y) = \frac{\|p_1(Y) - p_2(Y)\|^2 / (k-1)}{\tilde{\sigma}_1^2}$ que tem, sob H_2 distribuição $F_{(k-1, n-k)}$.

O numerador da estatística F é

$$\|p_1(Y) - p_2(Y)\|^2 = \sum_i n_i (\bar{y}_{i+} - \bar{y}_{++})^2.$$

representando uma medida da discrepância da média de cada grupo à média global.

Uma outra forma possível de calcular o numerador de F é usando o teorema de Pitágoras

$$\|p_1(Y) - p_2(Y)\|^2 = \|y - p_2(y)\|^2 - \|y - p_1(y)\|^2$$

Denotando os desvios para L_1 e para L_2 por $D_1 = \|y - p_1(y)\|^2$ e $D_2 = \|y - p_2(y)\|^2$, temos a seguinte fórmula para o teste F

$$F(y) = \frac{(D_2 - D_1)/(f_2 - f_1)}{D_1/f_1}$$

onde $f_1 = n - \dim L_1$ e $f_2 = n - \dim L_2$, são os graus de liberdade.

Outra fórmula comumente usada é

$$\|p_1(y) - p_2(y)\|^2 = \sum_i y_{i+}^2/n_i - \bar{y}_{++}/n,$$

que corresponde ao caminho tradicional na análise de variância. Usando a expressão acima temos a seguinte expressão para a estatística F

$$F(y) = \frac{\frac{1}{k-1} \sum_i n_i (\bar{y}_{i+} - \bar{y}_{++})^2}{\frac{1}{n-k} \sum_i \sum_j (y_{ij} - \bar{y}_{i+})^2}$$

Podemos então, interpretar a estatística F como a razão da variação entre amostras e dentro das amostras.

É comum representar os resultados de uma análise de variância através da seguinte tabela

Modelo	Desvio	g.l.	$\tilde{\sigma}_2^2$	F
I	D_1	f_1	$\tilde{\sigma}_1^2$	$F(y)$
<u>1</u>	D_2	f_2	$\tilde{\sigma}_2^2$	

Enquanto a tabela acima está num formato que usaremos para qualquer modelo, balanceado ou não, a tabela de análise de variância tradicional, adequada para experimentos balanceados é dada a seguir.

F. V.	Efeito	g.l.	Efeito médio	F
Entre am.	$D_2 - D_1$	$f_2 - f_1$	$(D_2 - D_1)/(f_2 - f_1)$	$F(y)$
Dentro am.	D_1	f_2	D_1/f_1	
Total	D_2	f_2	D_2/f_2	

Considere agora o caso onde L_2 foi rejeitado, indicando que os β_i 's não são iguais.

\Rightarrow testar hipóteses da forma $H_3 : \beta_i = \beta_r$ para i e r dados, sob L_1 .

O teste t para esta hipótese é dado por

$$t(Y) = \frac{\hat{\beta}_i - \hat{\beta}_r}{\tilde{\sigma}_1(1/n_i + 1/n_r)^{1/2}}$$

onde aqui usamos $Var(\hat{\beta}_i - \hat{\beta}_r) = \sigma^2(1/n_i + 1/n_r)$, que segue da independência entre $\hat{\beta}_i$ e $\hat{\beta}_r$. O teste tem $n - k$ graus de liberdade.

(EXEMPLO 3.3.4: Dados sobre bilirubina)

3.4 Parametrizações dos modelos lineares

3.4.1 Contrastes

O modelo a um fator apresentado na seção anterior nos fornece uma oportunidade para discutir algumas idéias sobre parametrizações. Até aqui o modelo foi parametrizado como

$\mu_{ij} = \beta_i, j = 1, \dots, n_i$ onde os β_i 's variam livremente. Mas, algumas vezes pode ser útil escrever o modelo como

$$\mu_{ij} = \alpha + \delta_i, j = 1, \dots, n_i,$$

sujeito a algumas restrições sobre os δ_i 's. A hipótese $\beta_1 = \dots = \beta_k$ no primeiro caso corresponde à $\delta_1 = \dots = \delta_k = 0$, no segundo. Similarmente, a hipótese $\beta_i = \beta_r$, corresponde à hipótese $\delta_i = \delta_r$.

Para lidar com o segundo caso, seja $\alpha \in \mathbb{R}$ arbitrário e escreva

$$\mu = \sum_i e_i \beta_i = \sum_i e_i (\beta_i - \alpha) + \sum_i e_i \alpha = \sum_i e_i \delta_i + \alpha \underline{1} \quad (2)$$

onde $\delta_i = \beta_i - \alpha$ e usamos os vetores dummy para I satisfazer a restrição $\sum_i e_i = \underline{1}$.

Existem muitas restrições diferentes sobre os δ_i 's que podem tornar a representação dada em (2) única.

Vejamos o caso que requer que as duas partes em (2) sejam ortogonais, isto é,

$(\sum_i e_i \delta_i) \cdot \underline{1} = 0$ ou $\sum_i (e_i \cdot \underline{1}) \delta_i = 0$ ou, equivalentemente, $\sum_i n_i \delta_i = 0$. Os δ_i 's que satisfazem esta última restrição são chamados contrastes.

O espaço das combinações lineares da forma

$\sum_i e_i \delta_i$ para δ_i satisfazendo $\sum_i n_i \delta_i = 0$ é um espaço linear.

Tal espaço, denotado por $L_1 \ominus L_2$ é, de fato, o complemento ortogonal para L_2 em L_1 .

Pela ortogonalidade das duas partes de (2), a projeção sobre L_1 é a soma das projeções sobre $L_1 \ominus L_2$ e L_2 . Vimos que

$p_2(y) = \bar{y}_{++}$ resultando em $\hat{\alpha} = \bar{y}_{++}$. Usando $\hat{\beta}_i = \bar{y}_{i+}$, obtemos

$$\hat{\delta}_i = \hat{\beta}_i - \hat{\alpha} = \bar{y}_{i+} - \bar{y}_{++}$$

que são os EMV's para a representação (2) sujeita à restrição $\sum_i n_i \delta_i = 0$.

Observe que para $n_1 = \dots = n_k$, o caso balanceado, a restrição anterior reduz-se a $\sum_i \delta_i = 0$.

É possível trabalhar com esta última condição em vez da anterior mesmo no caso não balanceado. Porém, a ortogonalidade que ajudou no fornecimento do estimador simples não é mais satisfeita.

3.4.2 A notação de Wilkinson e Rogers

Notação especial (Wilkinson e Rogers, 1973)

Princípio básico: um modelo linear

$L = \text{span}\{x_1, \dots, x_k\}$ onde x_1, \dots, x_k são vetores de \mathbb{R}^n é representado como $x_1 + \dots + x_k$.

Tal representação é chamada fórmula do modelo.

Mais precisamente, um conjunto finito $A \subseteq \mathbb{R}^n$ representa o espaço linear $\text{span}\{A\}$ e o operador “+” é definido por

$A + B = \text{span}\{A \cup B\} = \text{span}\{A\} + \text{span}\{B\}$ para A e B conjuntos dados. O operador “-” é definido por

$$A - B = \text{span}\{A \setminus B\}.$$

Para uma fórmula de modelo $x_1 + \dots + x_k$, a sequência na qual os termos aparecem não influencia o correspondente espaço linear, mas adotamos a convenção de que a sequência define a escolha da base para o modelo, de acordo com as seguintes regras. Se os x_j ’s ainda não formam uma base, uma base é selecionada da sequência x_1, \dots, x_k de acordo com o princípio de que se $x_j \in \text{span}\{x_1, \dots, x_{j-1}\}$, então x_j é excluído. Um x_j excluído por este processo costuma ser chamado de “aliased”.

O fenômeno onde um conjunto de vetores é linearmente dependente é algumas vezes chamado de multicolinearidade.

Se I denota um fator, adotamos a convenção de que o termo I em uma fórmula de modelo representa o conjunto de k vetores dummy e_1, \dots, e_k correspondendo ao fator I . A fórmula de modelo I representa, assim, o modelo a um fator.

Convenção: fatores são representados por letras maiúsculas tais como I , J e K enquanto que letras minúsculas como x , x_1 , etc., representam vetores.

O uso de um fator I em uma fórmula de modelo representa uma simplificação considerável comparada com a fórmula equivalente $e_1 + \dots + e_k$ pois evita a formação explícita dos vetores e_i 's.

Certos pacotes incluem automaticamente o vetor constante $\underline{1}$ na fórmula do modelo. Assim, por exemplo, a fórmula x representa na prática $\underline{1} + x$. Isto pode ser útil na prática mas leva resultados do tipo I , $I + \underline{1}$ e $I - \underline{1}$ todos representando o mesmo modelo. Assim, não adotaremos esta convenção.

Se I e J são dois fatores, definimos o novo fator $I.J$ como tendo níveis dados por cada combinação possível dos níveis de I e J . O fator $I.J$ é chamado de interação entre I e J . Também define-se o fator I/J como $I/J = I + I.J$.

O fator saturado é aquele com n níveis, uma para cada observação. O modelo saturado não tem uso prático real mas serve como uma referência conveniente correspondente ao maior modelo linear possível, ou em outras palavras, R^n .

Para o modelo a um fator (onde na prática podemos definir os fatores I e J), o modelo $I.J$ é o modelo saturado, assim como o modelo I/J . Porém estes modelos são representados por bases diferentes.

Deve-se sumariar a estrutura de um dado conjunto de modelos relacionados pelo que chamamos de diagrama do modelo. Por exemplo, o diagrama do modelo para submodelos do modelo a um fator (incluindo o modelo saturado) é

$$I + I/J \rightarrow \underline{1} + I \rightarrow \underline{1}$$

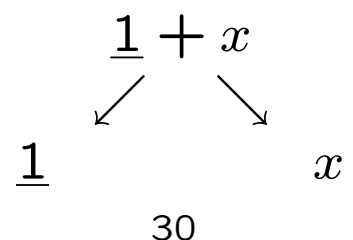
Aqui, uma seta entre dois modelos indica que o segundo é submodelo do primeiro. O menor modelo pode ser testado sob o maior, geralmente por um teste F tal que cada seta é associada com um teste.

O diagrama do modelo é particularmente útil para guardar traços das várias sequências possíveis de hipóteses que podem ser examinadas por um dado tipo de modelo.

Como cada seta corresponde geralmente a remover um termo na fórmula do modelo, podemos associar à cada seta no diagrama um efeito D_{termo} correspondendo ao termo removido. Por exemplo, no diagrama apresentado, $D_{I.J} = D_1$ e $D_I = D_2 - D_1$.

Em particular, o efeito dentro das amostras D_1 formalmente corresponde à remover o termo de interação $I.J$. Observe porém que o efeito associado com um termo dado, geralmente depende da sequência na qual os termos são testados tal que um diagrama de modelo correspondendo a diferentes ordens de remoção de termos produzirá outros valores para o efeito dos termos.

Modelo de regressão linear simples:



3.5 Teste para a homogeneidade das variâncias

Sejam Y_{ij} independentes com $Y_{ij} \sim N(\beta_i, \sigma^2)$, $i = 1, \dots, k$, $j = 1, \dots, n_i$.

$\Rightarrow \sigma^2$ é constante para todos os grupos.

Se tal suposição não é satisfeita temos um modelo da forma

$$Y_{ij} \sim N(\beta_i, \sigma_i^2)$$

\Rightarrow Para cada uma das k amostras temos um modelo simples de uma amostra (seção 3.1). Em particular, obtemos os seguintes estimadores para β_i e σ_i^2 :

$\hat{\beta}_i = \bar{y}_{i+}$ e $\tilde{\sigma}_i^2 = \frac{D_i}{n_i}$ onde $D_i = \sum_j (y_{ij} - \bar{y}_{i+})^2$ é o desvio para o i -ésimo grupo.

$$D_i \sim \sigma_i^2 \chi_{(n_i-1)}^2 = Ga(\sigma_i^2(n_i - 1), (n_i - 1)/2)$$

onde $Ga(\mu, \lambda)$ denota uma distribuição Gama com parâmetros μ e λ . Pela independência das amostras temos que D_1, \dots, D_k são independentes pois D_i somente depende da i -ésima amostra.

\Rightarrow Testar $\sigma_1^2 = \dots = \sigma_k^2$ baseado na distribuição dos D_i 's.

Generalizando um pouco mais o problema, consideraremos o caso onde D_1, \dots, D_k são independentes com $D_i \sim Ga(\alpha_i \lambda_i, \lambda_i)$ onde $\lambda_1, \dots, \lambda_k$ são conhecidos.

Teste da razão de verossimilhança para $H_0 : \alpha_1 = \dots = \alpha_k$.

A função de verossimilhança para $\alpha_1, \dots, \alpha_k$ é

$$L(\alpha_1, \dots, \alpha_k) = \prod_i \frac{1}{\alpha_i^{\lambda_i} \Gamma(\lambda_i)} d_i^{\lambda_i - 1} \exp\{-d_i/\alpha_i\}$$

onde d_1, \dots, d_k são os valores observados dos k desvios. Como $(\alpha_1, \dots, \alpha_k)$ varia em \mathbb{R}_+^k e L é um produto de funções cada uma envolvendo um α_i .

L pode ser maximizada, maximizando-se cada uma destas funções separadamente.

Assim, os EMV's para $\alpha_1, \dots, \alpha_k$ são $\hat{\alpha}_i = \frac{d_i}{\lambda_i}$, $i = 1, \dots, k$.

Sob H_0 , a verossimilhança para α , o valor comum de $\alpha_1, \dots, \alpha_k$ é

$$L(\alpha) = \alpha^{-\lambda_+} \exp^{-d_+/\alpha} \prod_i \frac{1}{\Gamma(\lambda_i)} d_i^{\lambda_i-1}$$

\Rightarrow o EMV de α é $\hat{\alpha} = \frac{d_+}{\lambda_+}$.

Seja $Q(d) = 2 \log\{L(\hat{\alpha}_1, \dots, \hat{\alpha}_k)/L(\hat{\alpha})\}$ a estatística de teste dada pelo log. da razão de verossimilhança.

$$Q(d) = 2 \log \frac{\prod_i \alpha_i^{-\lambda_i}}{\hat{\alpha}^{-\lambda_+}} = 2\{\lambda_+ \log \hat{\alpha} - \sum_i \lambda_i \log \hat{\alpha}_i\}$$

Voltando ao modelo da forma $D_i \sim Ga(\sigma_i^2 f_i, f_i/2)$ com $f_i = n_i - 1$ e $\alpha_i = 2\sigma_i^2$ temos

$$Q(d) = f_+ \log 2\tilde{\sigma}^2 - \sum_i f_i \log 2\tilde{\sigma}_i = f_+ \log \tilde{\sigma}^2 - \sum_i f_i \log \tilde{\sigma}_i^2$$

onde $\tilde{\sigma}_i^2 = d_i/f_i$ e $\tilde{\sigma}^2 = d_+/f_+$.

Observe que $\tilde{\sigma}^2$ é uma média ponderada de $\tilde{\sigma}_1^2, \dots, \tilde{\sigma}_k^2$ com pesos f_1, \dots, f_k . $\tilde{\sigma}^2$ é chamado estimador combinado “pooled”.

Bartlett (1937) mostrou que se definimos $B(d) = Q(d)/C$, com

$$C = 1 + \frac{1}{3(k-1)}\{\sum_i 1/f_i - 1/f_+\},$$

então, aproximadamente, $B(D) \sim \chi_{(k-1)}^2$, onde $D = (D_1, \dots, D_k)^T$ para $\min\{f_1, \dots, f_k\} \rightarrow \infty$.

Bartlett mostrou que a aproximação acima pode ser usada para $\min\{f_1, \dots, f_k\} \geq 2$. Observe que $C \rightarrow 1$.

De fato, espera-se que $Q(D) \sim \chi^2_{(k-1)}$ aproximadamente com base na teoria de grandes amostras pois $f_i = n_i - 1$ representa essencialmente tamanho amostral.

A contribuição de Bartlett foi calcular o fator de correção para valores de f_i relativamente pequenos.