Bootstrap e jackknife

Cristiano de Carvalho Santos
cristcarvalhosan@gmail.com

Departamento de Estatística,
Universidade Federal de Minas Gerais (UFMG)

Introdução

- O bootstrap foi introduzido por Efron em 1979, com mais desenvolvimentos nos anos seguintes.
- Métodos bootstrap são uma classe métodos Monte Carlo que estimam a distribuição de uma população por reamostragem e são frequentemente utilizados quando a distribuição da população alvo não é especificada e a amostra é a única informação disponível.
- Eforon deu o nome bootstrap porque ao usar o método parece estar puxando-se por seu próprio bootstrap.

- Métodos de reamostragem tratam uma amostra observada como uma população finita e amostras aleatórias são geradas dela para estimar características populacionais e fazer inferência sobre a população amostrada.

- Métodos Monte Carlo que envolvem amostragem de uma distribuição de probabilidade completamente especificada são algumas vezes chamados de bootstrap paramétrico.

Ideia do bootstrap

- A distribuição da população finita representada pela amostra pode ser considerada como uma pseudo-população com características semelhantes à da população verdadeira.
- Ao gerar repetidamente amostras aleatórias a partir dessa pseudo-população (reamostragem), a distribuição amostral de uma estatística pode ser estimada.
- Logo, propriedades de um estimador, como viés ou erro padrão, podem ser estimadas.

As estimativas de Bootstrap de uma distribuição de amostragem são análogas à ideia de estimativa de densidade:

- Um histograma não é a densidade, mas em um problema não paramétrico, pode ser visto como uma estimativa razoável da função de densidade.
- Temos métodos para gerar amostras aleatórias a partir de densidades completamente especificadas; bootstrap gera amostras aleatórias a partir da distribuição empírica da amostra.

Suponha que $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ é uma amostra aleatória observada de uma distribuição com fda $F(x)$. Se X^* é selecionado aleatoriamente de $\boldsymbol{x}$, então

$$P(X^* = x_i) = \frac{1}{n}, \quad i = 1, \ldots, n.$$

- Reamostragem gera uma amostra aleatória $X_1^*, \ldots, X_n^*$ por amostrar com reposição de $\boldsymbol{x}$, logo X^* são iid com distribuição Uniforme no conjunto $\{x_1, \ldots, x_n\}$.
- A fda empírica $F_n(x)$ é um estimador de $F(x)$ e também é a fda de X^*, pois reamostrar de $\boldsymbol{x}$ é equivalente a gerar da distribuição $F_n(x)$.

Seja $\tilde{\theta} = g(F_n)$ o valor do parâmetro quando F_n é a distribuição, $\hat{\theta} = s(\boldsymbol{x})$ uma estimativa de $\theta = g(F)$ obtida com a amostra observada e $\hat{\theta}^*$ uma estimativa obtida com uma amostra bootstrap $\boldsymbol{x}^*$.

Obs.: Frequentemente $\tilde{\theta}$ e $\hat{\theta}$ são iguais, mas podem ser diferentes. Por exemplo, se $\hat{\theta}$ for a média aparada dos dados e $\tilde{\theta}$ for a média da distribuição F_n.

Métodos bootstrap fazem um ou outra das grandes suposições a seguir:

A - A fda empírica F_n é uma boa aproximação da fda F_X, então a distribuição de $\hat{\theta}^*$ é similar a distribuição de $\hat{\theta}$.

B - A distribuição de $\hat{\theta}^* - \tilde{\theta}$ é similar a distribuição de $\hat{\theta} - \theta$.

- Em ambos os pressupostos, a tarefa de fazer inferências sobre θ se reduz a aprender sobre a distribuição bootstrap de $\hat{\theta}^*$.
- Às vezes, os aspectos relevantes da distribuição de bootstrap podem ser determinados matematicamente, mas, na maioria dos problemas não-triviais, a distribuição deve ser estimada usando métodos Monte Carlo.

A estimativa bootstrap da distribuição de $\hat{\theta}$ é obtido por:

1. Para cada réplica bootstrap, indexada por $b = 1, \ldots, B$:
   a) Gere uma amostra

$$\boldsymbol{x}^{*(b)} = (x_1^*, \ldots, x_n^*)^T$$

   por amostrar com reposição da amostra observada $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.

   b) Calcule $\hat{\theta}^{(b)}$ com a b-ésima amostra bootstrap.

2. A estimativa bootstrap de $F_{\hat{\theta}}$ é dada pela distribuição empírica de $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$.

Estimação do erro padrão

A estimativa bootstrap do erro padrão de um estimador $\hat{\theta}$ é o desvio padrão amostral das réplicas bootstrap $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$. Isto é,

$$\hat{se}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left[\hat{\theta}^{(b)} - \bar{\theta}\right]^2},$$

em que $\bar{\theta} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$.

Obs: De acordo com Efron e Tibshirani, o número de réplicas necessárias para boas estimativas do erro padrão não é grande, $B = 50$ é suficiente usualmente e raramente temos $B > 200$.

Exemplo

O conjunto de dados da escola de direito no pacote bootstrap é de Efron e Tibshirani. O quadro de dados contém LSAT (pontuação média na pontuação do teste de admissão na faculdade de direito) e GPA (nota média na metade da graduação) para 15 escolas de direito.

LSAT 576 635 558 578 666 580 555 661 651 605 653 575 545 572 594
GPA 339 330 281 303 344 307 300 343 336 313 312 274 276 288 296

- Este conjunto de dados é uma amostra aleatória do universo de 82 faculdades de direito (law82 - bootstrap).
- O objetivo é estimar a correlação entre as pontuações LSAT e GPA e calcular a estimativa de bootstrap do erro padrão da correlação da amostra.

Estimação do viés

O viés de um estimador $\hat{\theta}$ para θ é

$$Viés(\hat{\theta}) = E_F[\hat{\theta} - \theta] = E_F[s(\boldsymbol{X})] - g(F),$$

Uma estimativa bootstrap do víes é obtida ao substituir F por F_n e assim

$$\widehat{Viés}(\hat{\theta}) = E_{F_n}[s(\boldsymbol{X}^*)] - g(F_n) = \bar{\theta} - \tilde{\theta},$$

em que $\bar{\theta} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{(b)}$.

Exemplo

Os dados de Efron e Tibshirani contêm medidas de um certo hormônio na corrente sanguínea de oito indivíduos após o uso de um adesivo médico. O parâmetro de interesse é

$$\theta = \frac{E(new) - E(old)}{E(old) - E(placebo)}.$$

Se $|\theta| \leq 0,2$, isso indica bioequivalência dos adesivos antigo e novo. A estatística é $\bar{Y}/\bar{Z}$.

Desejamos calcular uma estimativa bootstrap de viés na estatística de razão de bioequivalência.

subject	placebo	oldpatch	newpatch	old-plac. z	new-old y
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719
mean:				6342	-452.3

Jackknife

- É um método de reamostragem proposto por Quenouille(1949) como uma técnica para redução de viés e por Tukey para estimar o erro padrão.

- No Jackknife, como em um tipo de validação cruzada, são consideradas subamostras em que cada x_i é omitido.

Seja $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ é uma amostra aleatória observada de uma distribuição com fda $F(x)$.

Definimos $\boldsymbol{x}_{[-i]} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)^T$ o subconjunto de $\boldsymbol{x}$ sem a i-ésima observação.

Seja $\hat{\theta} = T_n(\boldsymbol{x})$ e $\hat{\theta}_{[-i]} = T_n(\boldsymbol{x}_{[-i]})$, $i = 1, \ldots, n$.

Suponha que

- O parâmetro $\theta = g(F)$ é uma função da distribuição F;
- F_n é a fda empírica de uma amostra aleatória de F;
- A estimativa "plug-in" de θ é $\hat{\theta} = g(F_n)$.
- Um "plug-in" $\hat{\theta}$ é suave no sentido que pequenas mudanças nos dados correspondem a pequenas mudanças em $\hat{\theta}$.

Jackknife para estimar viés

Se $\hat{\theta}$ é uma estatística suave, então $\hat{\theta}_{[-i]} = g(F_{n-1}(x_{[-i]}))$ e a estimativa jackknife do viés é dada por

$$\widehat{Viés}_{jack}(\hat{\theta}) = (n-1)(\bar{\theta}_{[.]} - \hat{\theta}),$$

em que $\bar{\theta}_{[.]} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{[-i]}$ é a média das estimativas obtidas com as amostras com uma observação retirada.

O fator $n-1$ aparece para obter estimador jackknife não viesado para o viés do estimador plug-in da variância populacional.

Então, um estimador jackknife é dado por

$$\hat{\theta}_J = n\hat{\theta} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\theta}_{[-i]}.$$

Jackknife para estimar erro padrão

Uma estimativa jackknife do erro padrão é

$$\widehat{se}_{jack}(\hat{\theta}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{[-i]} - \bar{\theta}_{[.]}\right)^2},$$

para estatísticas suaves $\hat{\theta}$.

O fator $\frac{n-1}{n}$ faz com que $\hat{se}_{jack}$ seja um estimador não viciado do erro padrão da média.

Exemplo

Continuação do exemplo dos adesivos com hormônio e exemplo com a mediana no script!

Intervalos de confiança bootstrap

Existem várias abordagens para obter intervalos de confiança aproximados para o parâmetro de interesse. Entre eles, os intervalos de confiança:

- bootstrap normal padrão;
- bootstrap básico;
- bootstrap percentílico;
- bootstrap t.

O intervalo de confiança bootstrap normal padrão

Este intervalo de confiança possui uma abordagem simples, mas não necessariamente a melhor.

Se $\hat{\theta}$ é uma média amostral e o tamanho amostral é grande, então o Teorema Central do Limite implica que

$$Z = \frac{\hat{\theta} - E[\hat{\theta}]}{se(\hat{\theta})}$$

é aproximadamente normal padrão.

Logo, se $\hat{\theta}$ é um estimador não viesado para θ, um intervalo de confiança $100(1-\alpha)\%$ para θ é o intervalo

$$\hat{\theta} \pm z_{\alpha/2} se(\hat{\theta}),$$

em que $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$.

Este intervalo é simples, mas faz muitas suposições:

- A distribuição de $\hat{\theta}$ é normal ou $\hat{\theta}$ é a média amostral e o tamanho amostral é grande.
- $\hat{\theta}$ é um estimador não viciado de θ. O viés pode ser estimado e usado para centrar a distribuição de Z, mas o estimador é uma variável aleatória e a variável transformada não tem distribuição Normal.
- $se(\hat{\theta})$ é tratado como conhecido, mas é estimado.

O intervalo de confiança bootstrap básico

Este intervalo transforma a distribuição das réplicas do estimador por subtrair o valor observado da estatística.

Suponha que T é um estimador de θ e a_α tal que

$$P(T - \theta > a_\alpha) = 1 - \alpha \rightarrow P(T - a_\alpha > \theta) = 1 - \alpha.$$

Disso, o intervalo de confiança $100(1 - 2\alpha)\%$ é

$$(t - a_{1-\alpha}, \quad t - a_\alpha).$$

O percentil de ordem α de $\hat{\theta}^* - \hat{\theta}$ pode ser estimado por $\hat{b}_\alpha = \hat{\theta}_\alpha - \hat{\theta}$.

O limite superior do intervalo aproximado é dado por

$$\hat{\theta} - \hat{b}_\alpha = \hat{\theta} - (\hat{\theta}_\alpha - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}_{\alpha/2}$$

e, similarmente, o limite inferior do intervalo é dado por $2\hat{\theta} - \hat{\theta}_{1-\alpha/2}$.

Então, o intervalo de confiança $100(1-\alpha)\%$ é dado por

$$(2\hat{\theta} - \hat{\theta}_{1-\alpha/2}, \quad 2\hat{\theta} - \hat{\theta}_{\alpha/2}).$$

O intervalo de confiança bootstrap percentílico

- Esta abordagem utiliza a distribuição empírica das réplicas bootstrap como a distribuição de referência.

Suponha que $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$ são as réplicas bootstrap da estatística $\hat{\theta}$.

O intervalo é dado por

$$(\hat{\theta}_{\alpha/2}, \quad \hat{\theta}_{1-\alpha/2}),$$

em que $\hat{\theta}_{\alpha/2}$ é o percentil empírico calculado com a amostra $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$.

O intervalo de confiança bootstrap t

O bootstrap t não usa distribuição t-Student como referência, mas usa a distribuição amostral de uma estatística (*studentized*) gerada por reamostragem.

O intervalo é dado por

$$(\hat{\theta} - t^*_{1-\alpha/2}\hat{se}(\hat{\theta}), \quad \hat{\theta} - t^*_{\alpha/2}\hat{se}(\hat{\theta})),$$

em que $\hat{se}(\hat{\theta})$, $t^*_{1-\alpha/2}$ e $t^*_{\alpha/2}$ são calculados como a seguir.

1. Calcule a estatística observada $\hat{\theta}$.

2. Para cada réplica, indexada por $b = 1, \ldots, B$:

   a) Gere uma amostra

   $$\boldsymbol{x}^{*(b)} = (x_1^*, \ldots, x_n^*)^T$$

   por amostrar com reposição da amostra observada $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.

   b) Calcule $\hat{\theta}^{(b)}$ com a b-ésima amostra bootstrap.

   c) Calcule ou estime o erro padrão $\hat{se}(\hat{\theta}^{(b)})$. (Uma estimativa bootstrap por reamostrar da amostra atual $\boldsymbol{x}^{(b)}$).

   d) Calcule $t^{(b)} = \frac{\hat{\theta}^{(b)} - \hat{\theta}}{\hat{se}(\hat{\theta}^{(b)})}$.

3. Encontre os quantis $t^*_{1-\alpha/2}$ e $t^*_{\alpha/2}$ da amostra ordenada de $t^{(b)}$.

4. Calcule $\hat{se}(\hat{\theta})$ dado pelo desvio padrão das réplicas $\hat{\theta}^{(b)}$.

5. Calcule os limites de confiança dados por

$$(\hat{\theta} - t^*_{1-\alpha/2}\hat{se}(\hat{\theta}), \quad \hat{\theta} - t^*_{\alpha/2}\hat{se}(\hat{\theta})).$$

Uma desvantagem desta abordagem é fazer um bootstrap para cada réplica b com o intuito de estimar $\hat{se}(\hat{\theta}^{(b)})$. Ou seja, são B bootstraps dentro de um bootstrap!

Exemplo

Comparação dos intervalos de confiança para a correlação nos dados da faculdade de direito.

Melhores intervalos de confiança bootstrap

- O melhor intervalo de confiança do bootstrap é chamado BCa para "viés corrigido" e "ajustado para aceleração".
- Intervalos BCa são uma versão modificada de intervalos percentuais que têm melhores propriedades teóricas e melhor desempenho na prática.
- Para um intervalo de confiança de $100(1-\alpha)\%$, os quantis habituais $\alpha/2$ e $1-\alpha/2$ são ajustados por dois fatores: uma correção para viés e uma correção para assimetria.
- A correção de viés é denotada z_0 e o ajuste de assimetria ou "aceleração" é dado por a.

Um intervalo bootstrap BCa de confiança de $100(1-\alpha)\%$ é calculado por

$$\begin{aligned}
\alpha_1 &= \Phi^{-1}\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{a}(\hat{z}_0 + z_{\alpha/2})}\right), \\
\alpha_2 &= \Phi^{-1}\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1-\hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right),
\end{aligned}$$

em que $z_\alpha = \Phi^{-1}(\alpha)$,

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I\{\hat{\theta}^{(b)} < \hat{\theta}\}\right) \quad \text{e} \quad \hat{a} = \frac{\sum_{i=1}^{n}(\bar{\theta}_{[.]} - \hat{\theta}_{[-i]})^3}{6(\sum_{i=1}^{n}(\bar{\theta}_{[.]} - \hat{\theta}_{[-i]})^2)^{3/2}}.$$

Os limites sãos quantis empíricos das réplicas bootstrap e o intervalo BCa é

$$(\hat{\theta}_{\alpha_1}, \quad \hat{\theta}_{\alpha_2}).$$

Exemplo: Aplicação da metodologia no problema de bioequivalência no script.

Estruturas mais gerais de dados

- No mundo real, um mecanismo de probabilidade desconhecido P fornece um conjunto de dados $\boldsymbol{x}$ observado.
- Em aplicações específicas, precisamos definir a regra de construção dos dados com mais cuidado. O conjunto de dados $\boldsymbol{x}$ pode não ser mais um único vetor. Ele tem uma forma dependente da estrutura de dados, por exemplo, $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{y})$ no problema de duas amostras.

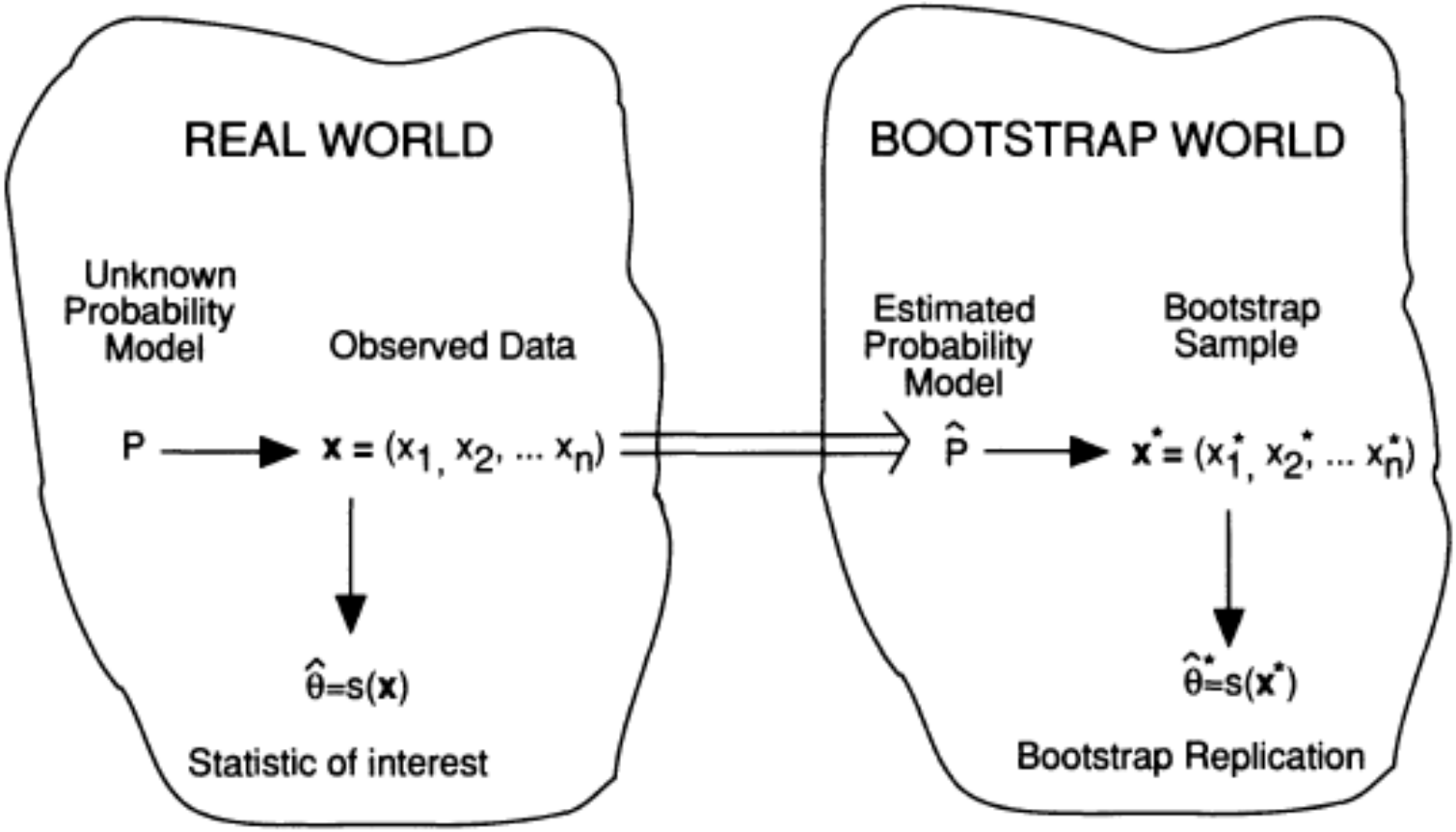

Figure 8.3. *Schematic diagram of the bootstrap applied to problems with a general data structure* $P \to \mathbf{x}$. *The crucial step* "$\Longrightarrow$" *produces an estimate* $\hat{P}$ *of the entire probability mechanism* P *from the observed data* $\mathbf{x}$. *The rest of the bootstrap picture is determined by the real world:* "$\hat{P} \to \mathbf{x}^*$" *is the same as* "$P \to \mathbf{x}$"; *the mapping from* $\mathbf{x}^* \to \hat{\theta}^*$, $s(\mathbf{x}^*)$, *is the same as the mapping from* $\mathbf{x} \to \hat{\theta}$, $s(\mathbf{x})$.

Dois problemas práticos surgem:

(1) Precisamos estimar todo o mecanismo de probabilidade P a partir dos dados observados $\boldsymbol{x}$. É fácil de fazer para a maioria das estruturas de dados familiares. Nenhuma prescrição geral é possível, mas soluções ad hoc bastante naturais estão disponíveis.

(2) Precisamos simular os dados de bootstrap de P de acordo com a estrutura de dados relevante. Este passo é conceitualmente direto, mas pode requerer algum cuidado na programação se a eficiência computacional for necessária.

Modelos de Regressão

O conjunto de dados $\boldsymbol{x}$ para um modelo de regressão linear consiste de n pontos

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n,$$

em que $\boldsymbol{x}_i = (\boldsymbol{c}_i, y_i)$, tal que $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{ip})$ é um vetor de covariáveis, enquanto que y_i é a variável resposta.

A suposição chave do modelo linear é que

$$\mu_i = E[Y_i|\boldsymbol{c}_i] = \boldsymbol{c}_i\boldsymbol{\beta} = \sum_{j=1}^{p} c_{ij}\beta_j.$$

O vetor de parâmetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ é desconhecido e objetivo usual a análise de regressão é fazer inferência sobre $\boldsymbol{\beta}$ a partir dos dados observados.

A estrutura de probabilidade do modelo linear é usualmente expressa como

$$y_i = \boldsymbol{c}_i\beta + \epsilon_i, \text{ para } i = 1, 2, \ldots, n.$$

Assumimos que os temos de erro ϵ_i são uma amostra aleatória de uma distribuição desconhecida F com esperança 0, isto é,

$$F \to (\epsilon_1, \ldots, \epsilon_n) \quad (E[\epsilon_i] = 0).$$

Note que

$$E[Y_i|\boldsymbol{c}_i] = E[\boldsymbol{c}_i\beta + \epsilon_i|\boldsymbol{c}_i] = \boldsymbol{c}_i\beta,$$

em que usamos o fato de que $E[\epsilon_i|\boldsymbol{c}_i] = E[\epsilon_i] = 0$, dado que ϵ_i é selecionado independentemente de $\boldsymbol{c}_i$.

Definimos o erro quadrático residual por

$$RSE(\boldsymbol{b}) = \sum_{i=1}^{n} (y_i - \boldsymbol{c}_i \boldsymbol{b})^2.$$

A estimativa de mínimos quadrados de $\boldsymbol{\beta}$ é o valor que minimiza $RSE(\boldsymbol{b})$.

Seja $\boldsymbol{C}$ com a i-ésima linha sendo $\boldsymbol{c}_i$ e seja $\boldsymbol{y}$ o vetor $(y_1, \ldots, y_n)^T$. Então a estimativa de mínimos quadrados é dada pela solução de

$$\boldsymbol{C}^T \boldsymbol{C} \hat{\boldsymbol{\beta}} = \boldsymbol{C}^T \boldsymbol{y}$$

que é dada por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{C}^T \boldsymbol{C})^{-1} \boldsymbol{C}^T \boldsymbol{y}.$$

O erro padrão de $\hat{\beta}_j$ é dado por

$$se(\hat{\beta}_j) = \sigma_F \sqrt{G^{jj}},$$

em que G^{jj} é o j-ésimo elemento da diagonal da matriz inversa $\boldsymbol{G}^{-1}$, tal que $\boldsymbol{G} = \boldsymbol{C}^T\boldsymbol{C}$ e $\sigma_F^2 = Var_F(\epsilon)$.

Na prática, σ_F^2 é estimado por

$$\hat{\sigma}_F^2 = \sum_{i=1}^{n}(y_i - \boldsymbol{c}_i\hat{\boldsymbol{\beta}})^2/n = RSE(\hat{\boldsymbol{\beta}})/n,$$

ou pela versão com viés corrigido dada por

$$\tilde{\sigma}_F^2 = \sum_{i=1}^{n}(y_i - \boldsymbol{c}_i\hat{\boldsymbol{\beta}})^2/(n-p) = RSE(\hat{\boldsymbol{\beta}})/(n-p).$$

Os correspondentes erros padrão estimados para os componentes de $\hat{\boldsymbol{\beta}}$ são

$$\hat{se}(\hat{\beta}_j) = \hat{\sigma}_F\sqrt{G^{jj}} \quad \text{ou} \quad \tilde{se}(\hat{\beta}_j) = \tilde{\sigma}_F\sqrt{G^{jj}}.$$

Aplicação do bootstrap

- Nenhum dos cálculos até agora requer o bootstrap.
- No entanto, uma análise de bootstrap para o modelo de regressão linear pode ser útil para assegurar que o bootstrap está dando respostas razoáveis.
- Podemos aplicar o bootstrap a modelos de regressão mais gerais que não têm solução matemática: onde a função de regressão é não linear nos parâmetros β, e onde usamos métodos de ajuste diferentes de mínimos quadrados.

O modelo de probabilidade $P \to \boldsymbol{x}$ para regressão linear tem duas componentes,

$$P = (\boldsymbol{\beta}, F),$$

em que F é a distribuição de probabilidade dos termos de erro.

Temos disponível $\hat{\beta}$, mas como podemos estimar F?

Se $\boldsymbol{\beta}$ é conhecido, sabemos que $\epsilon_i = y_i - \boldsymbol{c}_i\boldsymbol{\beta}$ para $i = 1, \ldots, n$. Então podemos calcular uma aproximação para os erros

$$\hat{\epsilon}_i = y_i - \boldsymbol{c}_i\hat{\boldsymbol{\beta}}, \quad \text{para} \quad i = 1, \ldots, n.$$

A estimativa para F é a distribuição empírica de $\hat{\epsilon}_i$ dada por

$$\hat{F} : \text{probabilidade } 1/n \text{ de sair } \hat{\epsilon}_i, \ i = 1, \ldots, n.$$

Com $\hat{P} = (\hat{\boldsymbol{\beta}}, \hat{F})$, sabemos como gerar os conjuntos de dados bootstrap para o modelo de regressão linear: $\hat{P} \to \boldsymbol{x}^*$.
Para gerar $\boldsymbol{x}^*$, primeiro selecionamos uma amostra bootstrap dos erros aleatórios,

$$\hat{F} \to (\epsilon_i^*, \ldots, \epsilon_n^*) = \boldsymbol{\epsilon}^*.$$

Então, as respostas bootstrap y_i^* são geradas de acordo com

$$y_i^* = \boldsymbol{c}_i \hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}_i^*, \quad \text{para} \quad i = 1, \ldots, n.$$

A estimativa de mínimos quadrados bootstrap é dada por

$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{C}^T \boldsymbol{C})^{-1} \boldsymbol{C}^T \boldsymbol{y}^*.$$

Neste caso, não precisamos de simulações de Monte Carlo para descobrir erros padrão de bootstrap,

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}^*) &= (\boldsymbol{C}^T\boldsymbol{C})^{-1}Var(\boldsymbol{y}^*)\boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{C})^{-1} \\ &= \hat{\sigma}_F^2(\boldsymbol{C}^T\boldsymbol{C})^{-1}, \end{aligned}$$

dado que $Var(\boldsymbol{y}^*) = \hat{\sigma}_F^2\boldsymbol{I}$ em que $\boldsymbol{I}$ é a matriz identidade. Então,

$$\hat{se}_\infty(\hat{\beta}_j) = se_{\hat{F}}(\hat{\beta}_j^*) = \hat{\sigma}_F\sqrt{G^{jj}},$$

em que ∞ acima representa $B = \infty$.

Em outras palavras, a estimativa bootstrap do erro padrão para β_j é igual a estimativa usual!

Bootstrap dos pares vs bootstrap dos resíduos

Existem dois diferentes caminha para implementar um bootstrap para o modelo de regressão:

1. O método dos pares considera $\boldsymbol{x}_i = (\boldsymbol{c}_i, y_i)$ de modo que um conjunto de dados de bootstrap $\boldsymbol{x}^*$ é da forma

$$\boldsymbol{x}^* = \{(\boldsymbol{c}_{i_1}, y_{i_1}), \ldots, (\boldsymbol{c}_{i_n}, y_{i_n})\}$$

   para $i_1, \ldots, i_n$ sendo uma amostra aleatória dos inteiros de 1 a n.

2. O método apresentado anteriormente pode ser chamado de bootstrap dos resíduos. Ele produz conjuntos de dados da forma

$$\boldsymbol{x}^* = \{(\boldsymbol{c}_1, \boldsymbol{c}_1\hat{\boldsymbol{\beta}} + \hat{\epsilon}_{i_1}), \ldots, (\boldsymbol{c}_n, \boldsymbol{c}_n\hat{\boldsymbol{\beta}} + \hat{\epsilon}_{i_n})\}.$$

Qual método de bootstrap é melhor?

A resposta depende de até que ponto acreditamos no modelo de regressão linear. Esse modelo assume que os erros tem a mesma distribuição F para qualquer valor de $\boldsymbol{c}_i$. Esta suposição pode falhar mesmo se a esperança $\mu_i = \boldsymbol{c}_i\beta$ esteja correta.

1. O bootstrap dos pares é menos sensível as suposições do modelo. A estimativa de erro padrão obtida por bootstrap dos pares dá respostas razoáveis, mesmo que as suposições do modelo de regressão estejam completamente erradas.

    - A única suposição por trás deste bootstrap é que os pares originais $\boldsymbol{x}_i = (\boldsymbol{c}_i, y_i)$ foram amostrados aleatoriamente a partir de alguma distribuição F, onde F é uma distribuição em vetores $(\boldsymbol{c}, y)$ com dimensão $(p + 1)$.

- Mesmo que as suposições do modelo de regressão estejam corretas, não é um desastre. Pode-se mostrar que as estimativas obtidas por esse método se aproximam daquelas dadas pelo bootstrap dos resíduos quando o número de pares n se torna grande.

2. O argumento inverso também pode ser feito. O modelo de regressão não precisa se manter perfeito para que os resíduos possam dar resultados razoáveis. Além disso, as diferenças nas distribuições de erro podem ser incorporadas no modelo, levando a uma versão mais apropriada dos resíduos de bootstrap.

Ponto importante: O bootstrap pode ser implementado de maneiras diferentes para o mesmo problema, dependendo de como o modelo de probabilidade é interpretado.

Exemplo: Os dados de sobrevivência de células

Um radiologista realizou um experimento envolvendo 14 placas bacterianas. As placas foram expostas a várias doses de radiação e a proporção das células sobreviventes foi medida. Doses maiores levam a menores proporções de sobrevivência, como seria de se esperar. O ponto de interrogação após a resposta para a placa 13 reflete alguma incerteza nesse resultado expresso pelo investigador.

plate number	dose (rads/100)	survive prop.	log.surv prop.
1	1.175	0.44000	-0.821
2	1.175	0.55000	-0.598
3	2.350	0.16000	-1.833
4	2.350	0.13000	-2.040
5	4.700	0.04000	-3.219
6	4.700	0.01960	-3.219
7	4.700	0.06120	-2.794
8	7.050	0.00500	-5.298
9	7.050	0.00320	-5.745
10	9.400	0.00110	-6.812
11	9.400	0.00015	-8.805
12	9.400	0.00019	-8.568
13	14.100	0.00700?	-4.962?
14	14.100	0.00006	-9.721

O investigador estava interessado em uma análise de regressão, com variável preditora

$$\text{dose}_i = z_i \quad i = 1, 2, \ldots, 14$$

e variável resposta

$$\log\left(\text{proporção de sobrevivência}_i\right) = y_i \quad i = 1, 2, \ldots, 14.$$

Dois modelos teóricos diferentes de dano por radiação estavam disponíveis, um dos quais previa uma regressão linear,

$$\mu_i = E[y_i|z_i] = \beta_1 z_i,$$

e outro com regressão quadrática,

$$\mu_i = E[y_i|z_i] = \beta_1 z_i + \beta_2 z_i^2.$$

Não existe intercepto por que era conhecido que com uma dose zero a proporção de sobrevivência era 1, logo $y = \log 1 = 0$.

Seja

$$MSR(\boldsymbol{b}) = \text{median}(y_i - \boldsymbol{c}_i\boldsymbol{b})^2.$$

A estimativa de minima mediana dos quadrados da regressão (LMS) para β é o valor que minimiza $MSR(\boldsymbol{b})$, isto é,

$$MSR(\hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{b}}(MSR(\boldsymbol{b})).$$

	$\hat{\beta}_1$	$(\widehat{\text{se}})$	$\hat{\beta}_2$	$(\widehat{\text{se}})$	$\hat{\beta}_2/\widehat{\text{se}}$
1. Least Squares, 14 plates	-1.05	(.159)	.0341	(.0143)	2.46
2. Least Squares, 13 plates	-0.86	(.094)	.0086	(.0091)	0.95
3. Least Median of Squares	-0.83	(.272)	.0114	(.0362)	0.32
4. (Resampling residuals)		(.141)		(.0160)	

Os erros padrão foram obtidos pelos métodos bootstrap.

- Os erros padrão na linha 3 são baseados em bootstrap dos pares com B $= 400$ replicações.
- As covariáveis nos dados de sobrevivência celular foram números fixos, estabelecidos pelo investigador: ela escolheu as doses

$$1.175, 1.175, 2.35, \ldots, 14.100$$

para ter um bom experimento para discriminar entre os modelos de sobrevivência de radiação linear e quadrática. Isso torna o bootstrap dos resíduos mais interessante.

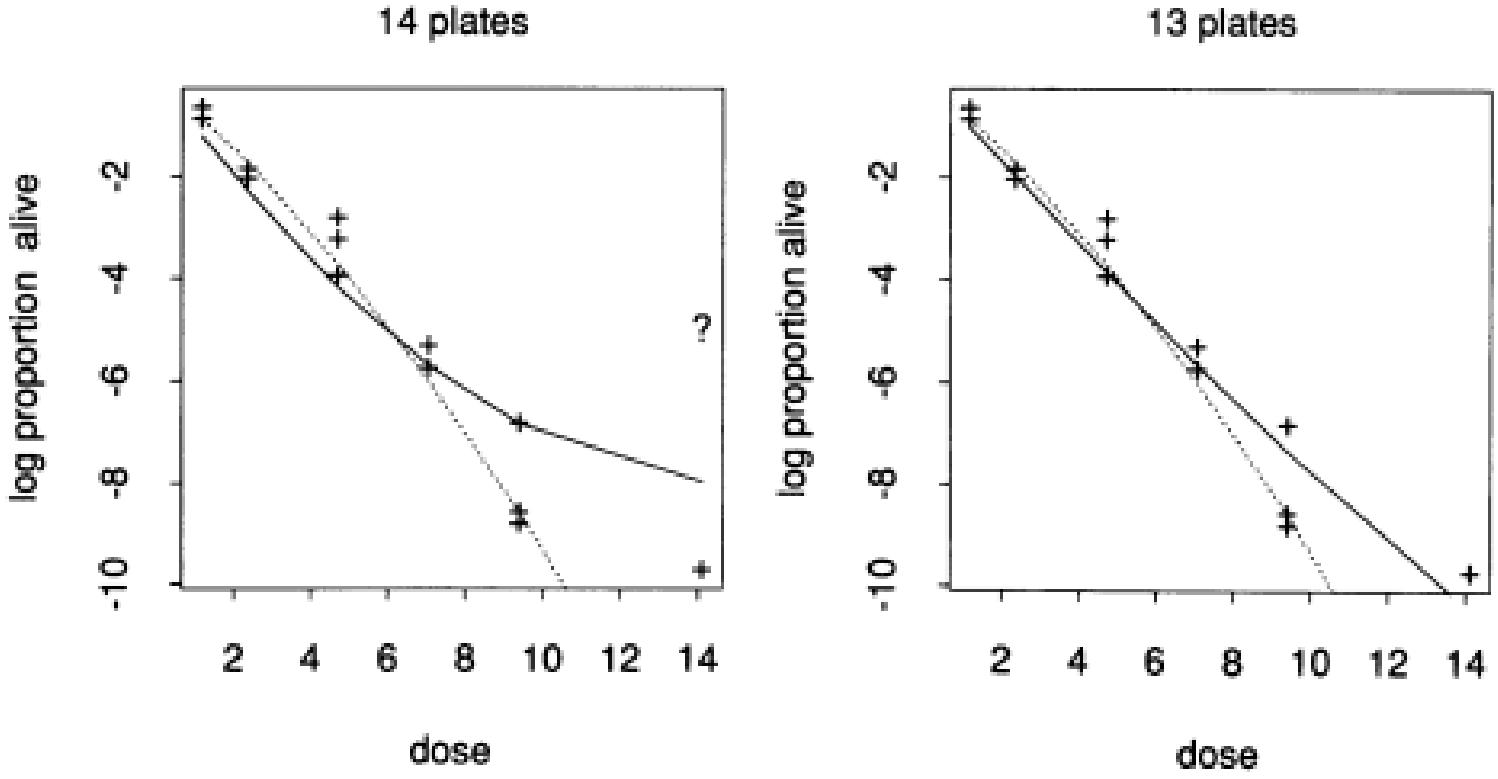

Figure 9.3. *Scatterplot of the cell survival data; solid line is the quadratic regression* $\hat{\beta}_1 z + \hat{\beta}_2 z^2$ *obtained by least-squares. Dashed line is quadratic regression fit by method of least median of squares (LMS). Left panel: all 14 plates; Right panel: thirteen plates, excluding the questionable result from plate 13. Plate 13, marked"?"in the left panel, has a large effect on the fitted least-squares curve. The questionable point has no effect on the LMS curve.*

Olhando para a Figura, podemos ver que a resposta y_i é mais dispersa para valores maiores de z. Como um modelo aproximadamente apropriado, assumiremos que os erros do modelo linear aumentam linearmente com a dose z. Isso equivale a

$$y_i = \boldsymbol{c}_i\boldsymbol{\beta} + z_i\epsilon_i \quad \text{para} \quad i = 1, \ldots, 14.$$

O vetor $\boldsymbol{\beta}$ foi estimado pelo LMS como $\hat{\boldsymbol{\beta}} = (-0.83, 0.0114)$. Então F foi estimada por $\hat{F}$, a distribuição empírica das quantidades

$$(y_i - \boldsymbol{c}_i\hat{\boldsymbol{\beta}})/z_i, \quad \text{para} \quad i = 1, \ldots, 14.$$

A linha 4 da Tabela 9.5 relata erros padrão de bootstrap para as estimativas de LMS obtidos com B = 200 replicações de bootstrap, utilizando o bootstrap dos resíduos.

Bootstrap paramétrico

- De fato, a amostragem de bootstrap pode ser realizada de forma paramétrica.

Quando os dados são modelados por uma distribuição paramétrica, isto é,

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} F(\boldsymbol{x}, \boldsymbol{\theta}),$$

uma outra estimativa de F pode ser obtida.

Suponha que os dados observados são usados para estimar $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$. No bootstrap paramétrico, Cada pseudo banco de dados $\boldsymbol{X}^*$ pode ser gerado por amostrar tal que

$$\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^* \overset{iid}{\sim} F(\boldsymbol{x}, \hat{\boldsymbol{\theta}}).$$

- Quando o modelo é conhecido ou acredita-se ser uma boa representação da realidade, o bootstrap paramétrico pode ser uma ferramenta poderosa:
  - permitindo inferência em situações de outra forma intratáveis;
  - produzindo intervalos de confiança muito mais precisos do que aqueles produzidos pela teoria assintótica padrão.
- É tentador usar um modelo conveniente, mas inadequado. Se o modelo não se encaixa bem no mecanismo que gera os dados, o bootstrap paramétrico pode levar a uma inferência errônea.
- Há ocasiões que poucas outras ferramentas inferenciais parecem viáveis.

Teste de hipóteses com o Bootstrap

- Os testes de bootstrap dão resultados semelhantes aos testes de permutação quando ambos estão disponíveis.
- Os testes de bootstrap são mais amplamente aplicáveis, embora menos precisos.

O problema de duas amostras

Observamos duas amostras aleatórias independentes $\boldsymbol{z}$ e $\boldsymbol{y}$ de possivelmente duas diferentes distribuições de probabilidade F e G,

$$\begin{aligned} F &\rightarrow \boldsymbol{z} = (z_1, \ldots, z_n) \text{ independente de} \\ G &\rightarrow \boldsymbol{y} = (y_1, \ldots, y_m) \end{aligned}$$

e desejamos testar a hipótese nula

$$H_0 : F = G.$$

Um teste de hipóteses é baseado em uma estatística de teste $t(\boldsymbol{x})$. Podemos considerar, por exemplo, $t(\boldsymbol{x}) = \bar{z} - \bar{y}$.

Podemos decidir sobre as hipóteses a partir de

$$\text{p-valor} = P_{H_0}(|t(\boldsymbol{x}^*)| \geq |t(\boldsymbol{x})|).$$

A quantidade $t(\boldsymbol{x})$ é o valor observado e a variável aleatória $\boldsymbol{x}^*$ tem uma distribuição especificada pela hipótese nula H_0, denotada por F_0.

Denotamos por $\boldsymbol{x}$ a amostra combinada de $\boldsymbol{z}$ e $\boldsymbol{y}$. A distribuição empírica de $\boldsymbol{x}$, denotada por $\hat{F}_0$, coloca probabilidade

$$1/(n+m)$$

para cada valor de $\boldsymbol{x}$.

Sob H_0, $\hat{F}_0$ fornece uma estimativa não-paramétrica da população comum que gerou $\boldsymbol{z}$ e $\boldsymbol{y}$.

Testes mais precisos podem ser obtidos através do uso de uma estatística estudentizada. Podemos utilizar

$$t(\boldsymbol{x}) = \frac{\bar{z} - \bar{y}}{\bar{\sigma}\sqrt{1/n + 1/m}},$$

em que

$$\bar{\sigma} = \left[\frac{\sum_{i=1}^{n} (z_i - \bar{z})^2 + \sum_{j=1}^{m} (y_j - \bar{y})^2}{n + m - 2} \right]^{1/2}.$$

Algoritmo

1. Gere B amostras de tamanho $n+m$ com reposição de $\boldsymbol{x}$. Para cada uma das amostras, denote as primeiras n observações por $\boldsymbol{z}^*$ e as m observações restantes por $\boldsymbol{y}*$.

2. Avalie $t(\cdot)$ em cada reamostra, isto é, $t(\boldsymbol{x}^{*b})$.

3. Calcule

$$\widehat{\text{p-valor}}_{boot} = \sum_{b=1}^{B} I\{|t(\boldsymbol{x}^{*b})| \geq |t_{obs}|\}/B,$$

em que $t_{obs} = t(\boldsymbol{x})$ é o valor observado da estatística de teste.

O algoritmo acima testa a hipótese nula de que as duas populações são idênticas, ou seja, $F = G$. E se quiséssemos testar apenas se suas médias eram iguais?

Se não estivermos dispostos a assumir que as variâncias nas duas populações são iguais, poderíamos basear o teste em

$$t(\boldsymbol{x}) = \frac{\bar{z} - \bar{y}}{\sqrt{\bar{\sigma}_1^2/n + \bar{\sigma}_2^2/m}},$$

em que

$$\bar{\sigma}_1 = \sum_{i=1}^{n} (z_i - \bar{z})^2/(n-1) \quad \text{e} \quad \bar{\sigma}_2 = \sum_{j=1}^{m} (y_j - \bar{y})^2/(m-1).$$

- A suposição de variância igual é atraente para o teste t porque simplifica a forma da distribuição da estatística de teste.
- Mas ao considerar um teste de hipótese de bootstrap para comparar as duas médias, não há razão convincente para assumir variações iguais e, portanto, não fazemos essa suposição.

Para prosseguir, precisamos de estimativas de F e G que utilizem apenas a suposição de uma média comum. Seja $\bar{x}$ a média da amostra combinada, podemos

1. transladar ambas as amostras de modo que tenham a média $\bar{x}$;
2. reamostrar de cada população separadamente.

Algoritmo

1. Faça $\hat{F}$ colocando igual probabilidade nos pontos

$$\tilde{z}_i = z_i - \bar{z} + \bar{x},$$

para $i = 1, \ldots, n$ e $\hat{G}$ colocando igual probabilidade nos pontos

$$\tilde{y}_i = y_i - \bar{y} + \bar{x},$$

para $i = 1, \ldots, m$, em que $\bar{z}$ e $\bar{y}$ são as médias dos grupos e $\bar{x}$ é a média da amostra combinada.

2. Gere B conjuntos de dados $(\boldsymbol{z}^*, \boldsymbol{y}^*)$ em que $\boldsymbol{z}^*$ é amostrado com reposição de $\tilde{z}_1, \ldots, \tilde{z}_n$ e $\boldsymbol{y}^*$ é amostrado com reposição de $\tilde{y}_1, \ldots, \tilde{y}_m$.

3. Para cada banco de dados calcule

$$t(\boldsymbol{x}) = \frac{\bar{z}^* - \bar{y}^*}{\sqrt{\bar{\sigma}_1^{*2}/n + \bar{\sigma}_2^{*2}/m}},$$

4. Calcule

$$\widehat{\text{p-valor}}_{boot} = \sum_{b=1}^{B} I\{|t(\boldsymbol{x}^{*b})| \geq |t_{obs}|\}/B,$$

em que $t_{obs} = t(\boldsymbol{x})$ é o valor observado da estatística de teste.

Exemplo

Simulação no script!

O problema com uma amostra

Suponha que observamos uma amostra $\boldsymbol{z}$ com distribuição de probabilidade F, isto é

$$F \quad \rightarrow \quad \boldsymbol{z} = (z_1, \ldots, z_n)$$

e desejamos testar a hipótese nula

$$H_0 : \mu_Z = \mu_0.$$

Um bootstrap pode ser utilizado considerando a estatística de teste

$$t(\boldsymbol{z}) = \frac{\bar{z} - \mu_0}{\bar{\sigma}/\sqrt{n}}.$$

Qual a distribuição da estatística de teste sob H_0?

- Necessitamos de uma distribuição $\hat{F}$ que estima da distribuição F sob H_0.
- Observe primeiro que a distribuição empírica $\hat{F}_n$ não é apropriada estimar para F porque não obedece a H_0. Ou seja, a média de F não é igual ao valor nulo de μ_0.

Uma maneira simples é transladar a distribuição empírica $\hat{F}_n$ para que tenha a média desejada. Em outras palavras, usamos como nossa distribuição nula estimada a distribuição empírica nos valores

$$\tilde{z}_i = z_i - \bar{z} + \mu_0, \quad \text{para} \quad i = 1, \ldots, n.$$

Então, amostramos

$$\tilde{z}_1^*, \ldots, \tilde{z}_n^*$$

com reposição de $\tilde{z}_1, \ldots, \tilde{z}_n$ e para cada amostra bootstrap calculamos a estatística

$$t(\boldsymbol{z}^*) = \frac{\bar{\tilde{z}}^* - \mu_0}{\bar{\tilde{\sigma}}^*/\sqrt{n}}.$$

Logo,

$$\widehat{\text{p-valor}}_{boot} = \sum_{b=1}^{B} I\{|t(\boldsymbol{z}^{*b})| \geq |t_{obs}|\}/B.$$

Existe uma maneira diferente, mas equivalente, de fazer um bootstrap no problema de uma amostra.

Amostramos com substituição dos dados originais (não transladados)

$$z_1, \ldots, z_n$$

e calculamos a estatística

$$t(\boldsymbol{z}^*) = \frac{\bar{z}^* - \bar{z}}{\sigma^*/\sqrt{n}},$$

em que σ^* é o desvio padrão da reamostra.

Esta estatística é igual a anterior pois

$$\bar{\tilde{z}}^* - \mu_0 = (\bar{z}^* - \bar{z} + \mu_0) - \mu_0 = \bar{z}^* - \bar{z}$$

e os desvios padrão também são iguais.

Exemplo

Simulação no script!

Testes de permutação

- Testes de permutação são baseados em reamostragem, mas as amostras são geradas sem reposição.
- Podem ser aplicados para realizar testes não paramétricos de igualdade de distribuições, independência, entre outros.

Igualdade de distribuições

Suponha duas amostras aleatórias independentes $\boldsymbol{z}$ e $\boldsymbol{y}$ de possivelmente duas diferentes distribuições de probabilidade F e G,

$$\begin{aligned} F &\rightarrow \boldsymbol{z} = (z_1, \ldots, z_n) \text{ independente de} \\ G &\rightarrow \boldsymbol{y} = (y_1, \ldots, y_m) \end{aligned}$$

Seja $\boldsymbol{x}$ a amostra agrupada

$$\boldsymbol{x} = (z_1, \ldots, z_n, y_1, \ldots, y_n),$$

que é indexada por

$$\boldsymbol{v} = \{1, \ldots, n, n+1, \ldots, n+m\} = \{1, \ldots, N\}.$$

Seja $\boldsymbol{X}^* = (\boldsymbol{Z}^*, \boldsymbol{Y}^*)$ representando uma partição da amostra agrupada $\boldsymbol{X}$, em que $\boldsymbol{Z}^*$ tem n elementos e $\boldsymbol{Y}^*$ tem $m = N - n$ elementos.

Então, $\boldsymbol{X}^*$ é uma permutação π dos inteiros $\boldsymbol{v}$, em que $z_i^* = z_{\pi(i)}$.

O número de possíveis partições é igual a $\binom{N}{n}$ e sob

$$H_0 : F = G,$$

uma amostra aleatória $\boldsymbol{Z}^*$ tem probabilidade

$$\frac{1}{\binom{N}{n}} = \frac{n!m!}{N!}$$

para quaisquer valores possíveis. Isto é, sob H_0 todas as permutações são igualmente prováveis.

- Os grupos podem ser comparados de várias maneiras. Por exemplo, com médias amostrais, medianas ou médias aparadas.
- Mais geralmente, pode-se perguntar se as distribuições das duas variáveis diferem e comparar os grupos por qualquer estatística que mede a distância entre duas amostras.

Se

$$\hat{\theta}(\boldsymbol{Z}, \boldsymbol{Y}) = \hat{\theta}(\boldsymbol{X}, \boldsymbol{v})$$

é uma estatística, então a distribuição de permutação de $\hat{\theta}^*$ é a distribuição de replicações

$$\{\hat{\theta}^*\} = \left\{ \hat{\theta}(\boldsymbol{X}, \pi_j(\boldsymbol{v})),\ j = 1, \ldots, \binom{N}{n} \right\}.$$

Assim,

$$\text{p-valor} = P(|\hat{\theta}^*| \geq |\hat{\theta}|) = \binom{N}{n}^{-1} \sum_{j=1}^{\binom{N}{n}} I\{|\hat{\theta}^{(j)}| \geq |\hat{\theta}|\},$$

onde $\hat{\theta}$ é o valor calculado com a amostra observada.

Obs: O p-valor pode ser calculado de maneira similar para um teste unilateral a esquerda ou bilateral.

- Na prática, a menos que o tamanho amostral seja muito pequeno, avaliar a estatística de teste para todas as permutações é computacionalmente intensivo.

- Um teste de permutação aproximado é implementado ao amostrar aleatoriamente um grande número de amostras sem reposição.

Algoritmo para teste de permutação aproximado

1. Calcule o valor observado $\hat{\theta}(\boldsymbol{Z}, \boldsymbol{Y}) = \hat{\theta}(\boldsymbol{X}, \boldsymbol{v})$ para a estatística de teste.

2. Para cada réplica, indexada por $b = 1, \ldots, B$:

   a) Gere uma permutação aleatória $\pi_b = \pi(\boldsymbol{v})$.
   b) Calcule a estatística $\hat{\theta}^{(b)} = \hat{\theta}(\boldsymbol{X}, \pi_b)$

3. Se grandes valores de $\hat{\theta}$ dão suporte a hipótese alternativa, calcule

$$\widehat{\text{p-valor}} = \frac{1 + \sum_{j=1}^{B} I\{\hat{\theta}^{(b)} \geq \hat{\theta}\}}{B + 1}.$$

4. Rejeite H_0 se $\widehat{\text{p-valor}} \leq \alpha$.

Exemplo

São registrados pesos em gramas, para seis grupos de pintinhos recém-nascidos alimentados com suplementos diferentes. Existem seis tipos de suplementos alimentares. Sugere-se que os grupos soja e linhaça podem ser semelhantes. A distribuição de pesos para esses dois grupos é comparada.

No script!

Teste de independência

Uma teste de independência de Z e Y dado por

$$H_0 : F_{ZY} = F_Z F_Y \quad \text{vs} \quad H_1 : F_{ZY} \neq F_Z F_Y$$

pode ser implementado como um teste de permutação?

- Se X e Y estão correlacionados/associados, eles são dependentes
- A hipótese nula é $H_0 : \rho = 0$ onde $\rho = cor(Z, Y)$
- Diferentes definições de ρ medem diferentes tipos de associação.

Como podemos usar um teste de permutação para responder a essa pergunta?

Seja $\boldsymbol{v} = (v_1, \ldots, v_n)$ o vetor de permutação que contém os inteiros $\{1, \ldots, n\}$ em alguma ordem. Esse vetor estará associado a ordenação de y_i.

Existem $n!$ possíveis vetores $\boldsymbol{v}$.

Se $H_0 : \rho = 0$ é verdadeira, então reordenar y_i não afetará a correlação.

Sob H_0, o vetor $\boldsymbol{v}$ em probabilidade $1/n!$ de assumir cada um dos $n!$ possíveis resultados.

Para o teste de $H_0 : \rho = 0$, temos que

$$\text{p-valor} = \frac{\sum_{j=1}^{n!} I\{|\hat{\rho}^{(j)}| \geq |\hat{\rho}|\}}{n!}.$$

Quando $n!$ é muito grande utilizamos a aproximação Monte Carlo.