

Une méthode bootstrap pratique pour tester des hypothèses à partir de données d'enquête

Jean-François Beaumont et ***Cynthia Bocci***

Jean-Francois.Beaumont@statcan.ca

Statistical Society of Ottawa Conference

29 novembre 2007

Motivation

- La méthode bootstrap est de plus en plus utilisée dans la pratique
 - Essentiellement, pour l'estimation de la variance de paramètres d'intérêt
 - Poids bootstrap sont fournis sur le fichier de données
- **Simple pour les analystes**
- Analystes préfèrent souvent utiliser des logiciels classiques comme SAS
- N'existe pas de méthodologie bootstrap pour tester des hypothèses dans les enquêtes

Contenu de la présentation

- Mise en contexte
- Méthode bootstrap
- Étude par simulations
- Conclusion

Mise en contexte

- Modèle d'analyse m : $F(\mathbf{y}_U | \mathbf{X}_U; \boldsymbol{\beta}, \boldsymbol{\theta})$
 - $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus du modèle et d'intérêt
- Supposons qu'on veuille tester l'hypothèse

$$H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c} \quad \text{vs} \quad H_1 : \mathbf{H}\boldsymbol{\beta} \neq \mathbf{c}$$

- Si la population U pouvait être observée:
 - On utiliserait une statistique $t(U; \mathbf{c})$ qui obéit à une distribution connue sous H_0

Mise en contexte

- On considère des statistiques de la forme:

$$t(U; \mathbf{c}) = \left(\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c} \right)' \{ \mathbf{A}(U) \}^{-1} \left(\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c} \right)$$

- Par exemple, si on a le modèle linéaire:

$$\mathbf{E}_m(y_k | \mathbf{X}_U) = \mathbf{x}'_k \boldsymbol{\beta} \quad \text{et} \quad \mathbf{V}_m(y_k | \mathbf{X}_U) = \theta$$

ou

$$\begin{array}{ll} 1) & \mathbf{A}(U) = \hat{\mathbf{V}}_m(\mathbf{H}\hat{\boldsymbol{\beta}}_U) \quad \Longrightarrow \quad t(U; \mathbf{c}) \square \chi_Q^2 \\ 2) & \mathbf{A}(U) = Q \hat{\mathbf{V}}_m(\mathbf{H}\hat{\boldsymbol{\beta}}_U) \quad \Longrightarrow \quad t(U; \mathbf{c}) \square F_{Q, N-r} \end{array}$$

Mise en contexte

- Comme on n'a qu'un échantillon s de U
 - On peut utiliser $t(s; \mathbf{c})$ au lieu de $t(U; \mathbf{c})$
 - On utilise plutôt:

$$\hat{t}(s, \mathbf{w}_s; \mathbf{c}) = \left(\mathbf{H} \hat{\boldsymbol{\beta}}_{ws} - \mathbf{c} \right)' \left\{ \hat{\mathbf{A}}(s, \mathbf{w}_s) \right\}^{-1} \left(\mathbf{H} \hat{\boldsymbol{\beta}}_{ws} - \mathbf{c} \right)$$

- Typiquement, les poids sont normalisés de telle sorte que leur somme = n
- Problème: $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ obéit à ??

Mise en contexte

- Rao et Scott (1981) et Fay (1985):
 - Modifie $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ de telle sorte que la statistique modifiée obéit à une distribution connue sous l'hypothèse nulle
- Bootstrap:
 - Approxime la distribution de $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ sous l'hypothèse nulle en utilisant les poids bootstrap

Méthode bootstrap

- Recette:

- Obtenir les poids bootstrap (normalisés) \mathbf{w}_s^{*b}
- Calculer $\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, $b = 1, \dots, B$
- En supposant qu'on rejette pour de grandes valeurs de $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$, on obtient le seuil observé (p-value):

$$\frac{\#\left\{\hat{t}(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\boldsymbol{\beta}}_{ws}) > \hat{t}(s, \mathbf{w}_s; \mathbf{c})\right\}}{B}$$

Exemple

- Échantillonnage aléatoire simple avec remise et un seul paramètre d'intérêt:

– Hypothèse nulle: $H_0 : \beta = c$ vs $H_1 : \beta \neq c$

– Statistique: $\hat{t}(s, \mathbf{w}_s; c) = \sqrt{n} \left(\hat{\beta}_{ws} - c \right)^2 / \hat{\theta}_{ws}$

– Statistique bootstrap:

$$\hat{t}(s, \mathbf{w}_s^{*b}; \hat{\beta}_{ws}) = \sqrt{n} \left(\hat{\beta}_{w^*b_s} - \hat{\beta}_{ws} \right)^2 / \hat{\theta}_{w^*b_s}$$

Méthode bootstrap

- Hypothèses principales:

1) $\sqrt{n}(\hat{\boldsymbol{\beta}}_{ws} - \boldsymbol{\beta}) \rightarrow N(\mathbf{0}, \mathbf{V})$ sous mp

2) $\sqrt{n}(\hat{\boldsymbol{\beta}}_{w_s^*} - \hat{\boldsymbol{\beta}}_{ws}) \rightarrow N(\mathbf{0}, \hat{\mathbf{V}})$ sous *

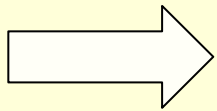
3) $\mathbf{V}_*(\hat{\boldsymbol{\beta}}_{w_s^*})$ est convergent pour $\mathbf{V}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$ sous mp

4) $\hat{\mathbf{A}}(s, \mathbf{w}_s^*)$ est convergent pour $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ sous *

Méthode bootstrap

- Note:

$$\begin{aligned} \mathbf{V}_{mp}(\hat{\boldsymbol{\beta}}_{ws}) &= \mathbf{E}_m \mathbf{V}_p(\hat{\boldsymbol{\beta}}_{ws}) + \mathbf{V}_m \mathbf{E}_p(\hat{\boldsymbol{\beta}}_{ws}) \\ &\approx \mathbf{E}_m \mathbf{V}_p(\hat{\boldsymbol{\beta}}_{ws}), \text{ si } n/N \text{ est négligeable} \end{aligned}$$



La variabilité due au modèle est négligeable

- Idéalement, on aimerait que:

5) La distribution asymptotique de $\hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta})$ sous mp ne dépende pas de $\boldsymbol{\beta}$ (et même de $\boldsymbol{\theta}$)

Étude par simulations

- Populations:
 - $N = 10\ 000$
 - Modèle ANOVA à un facteur et 5 niveaux
 - Hypothèse nulle: $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
- Plan de sondage:
 - 1000 échantillons stratifiés de taille 100
 - 2 scénarios de stratification: INFORMATIF et NON-INFORMATIF
- 500 poids bootstrap / échantillon sélectionné

Étude par simulations

- Méthodes:
 - Méthodes naïves:
 - Version pondérée: $\hat{t}(\mathbf{y}_s, \mathbf{w}_s; \mathbf{c}) \rightarrow F_{Q, n-r}$
 - Version non pondérée: $\hat{t}(\mathbf{y}_s, \mathbf{1}_s; \mathbf{c}) \rightarrow F_{Q, n-r}$
 - Invalides si l'échantillonnage est informatif
 - Version non pondérée est valide si l'échantillonnage n'est pas informatif
 - Wald, Bonferroni, Rao-Scott
 - Bootstrap

Taux de rejet de H_0 :

Échantillonnage informatif

Méthode	H_0 vraie	H_0 fausse
Naïve non pondérée	100.0	100.0
Naïve pondérée	0.5	5.0
Wald	16.0	38.1
Bonferroni	12.9	33.3
Rao-Scott	7.7	21.5
Bootstrap	6.9	20.9

Taux de rejet de H_0 :

Échantillonnage non informatif

Méthode	H_0 vraie	H_0 fausse
Naïve non pondérée	4.8	11.7
Naïve pondérée	12.6	23.8
Wald	8.8	19.6
Bonferroni	6.1	16.0
Rao-Scott	5.8	12.5
Bootstrap	3.8	9.0

Conclusion

- Méthode bootstrap est simple
 - Utilise des statistiques “model-based”
 - Pas besoin de logiciels spécialisés pour effectuer les tests
 - Facile pour les utilisateurs
- Performance de la méthode
 - Conduit à des tests valides (sous l’hypothèse nulle)
 - Perte légère de puissance dans certains cas

Thanks - Merci

For more
information
please contact

Pour plus
d'information,
veuillez
contacter

Jean-François Beaumont

Jean-Francois.Beaumont@statcan.ca

Cynthia Bocci

Cynthia.Bocci@statcan.ca