TP Statistiques 5

Juhyun Park, Angie Pineda, Nicolas Brunel

2 mai 2022

Normalité asymptotique de l'EMV et l'intervalle de confidence

On sait que la distribution asymptotique de l'estimateur du maximum de vraisemblance est

$$\hat{\theta} \sim \mathcal{N}(\theta, I_n(\theta)^{-1}),$$

où $I_n(\theta)$ est l'information de Fisher du modèle. Celle-ci fournit une base pour l'inférence statistique, telle que la construction d'un intervalle de confiance.

1. Soit une loi Normale avec $\mu = 2$ et $\sigma = 1$. Simuler un échantillon i.i.d de taille $n = 10$. Trouver des estimateurs de $\theta = (\mu, \sigma)$. Donner l'intervalle de confiance à 95% théorique pour μ. Construire des intervalles de confiance à 95% pour μ par la normalité asymptotique.
2. Répéter la simulation $N = 100$ fois pour construire les intervalles de confiance théoriques et asymptotiques. Combien de fois ces intervalles contiennent-ils le vrai paramètre? Confirment ils la couverture de 95%? Tracer l'histogramme des valeurs du maximum de vraisemblance. La distribution empirique de l'estimateur se rapproche-t-elle de la distribution asymptotique?
3. Répéter pour les tailles d'échantillon croissantes $n = 20, 50, 100$. Calculer la couverture empirique des intervalles de confiance. Quelle est l'influence de la taille de l'échantillon n sur l'inférence asymptotique?

Estimation de la covariance asymptotique

L'évaluation de la distribution asymptotique du maximum de vraisemblance avec l'aide de l'information de Fisher soulève plusieurs de questions. L'évaluation directe nécessite de calculer le hessien et son espérance analytiquement en différenciant l'opposé de la log-vraisemblance deux fois par rapport aux paramètres (pour obtenir la matrice complète), puis d'inverser explicitement la matrice, et enfin de remplacer θ par $\hat{\theta}$.

Cependant, il n'est pas toujours possible d'évaluer ces quantités de manière analytique, ou cela peut prendre trop de temps. En outre, Efron et Hinkley (1978, Biometrika) ont trouvé qu'il est préférable de travailler avec la matrice d'information "observée" plutôt que l'information "attendue" (c'est-à-dire l'information de Fisher qui est une espérance).

En guise de compromis, ces considérations conduisent les chercheurs à utiliser $I_O(\hat{\theta})^{-1} = -H(\hat{\theta})^{-1}$ comme matrice de variance en pratique, avec l'élément (i, j) de $H(\hat{\theta})$ obtenu en utilisant la différenciation finie, i.e.

$$\frac{\partial^2 \ell_n(\hat{\theta})}{\partial\theta_i \partial\theta_j} \approx \frac{\ell_n(\hat{\theta} + \epsilon_i 1_i + \epsilon_j 1_j) - \ell_n(\hat{\theta} + \epsilon_i 1_i) - \ell_n(\hat{\theta} + \epsilon_j 1_j) + \ell_n(\hat{\theta})}{\epsilon_i \epsilon_j},$$

ce qui est une bonne approximation lorsque $\{\epsilon_i\} \downarrow 0$. Ici $1_i = (0, \ldots, 0, 1, \ldots, 0)$ avec le 1 dans l'entrée i. Notez que cela ne nécessite que 3 nouvelles évaluations de log-vraisemblance par dérivée seconde.

4. Pour un échantillon de la loi Weibull avec les paramètres $\theta = (a, b) = (1.5, 3)$ de taille $n = 10$, estimer la covariance asymptotique d'estimateur. En utilisant cela, donner un intervalle de confiance asymptotique à 95% pour a et b. La vraie valeur est-elle incluse dans ces intervalles?

5. Répéter le calcul des intervalles de confiance $N = 100$ fois et donner la couverture empirique de l'intervalle de confiance. Est-ce proche de la valeur théorique?

6. Répéter pour les tailles d'échantillon croissantes $n = 20, 50, 100$. Calculer la couverture empirique des intervalles de confiance. Quelle est l'influence de la taille de l'échantillon n sur l'inférence asymptotique?

Méthode delta

La plupart des propriétés ci-dessus sont conservées lorsque nous sommes intéressés par l'estimation de $\phi = g(\theta)$. Ici, nous examinerons les problèmes numériques associés à cette étape de l'analyse.

La distribution asymptotique de $\hat{\phi}$ est donnée par

$$\hat{\phi} \sim \mathcal{N}(\phi, (\bigtriangledown_\theta g)^T I_n(\theta)^{-1} (\bigtriangledown_\theta g)).$$

Il est préférable de remplacer $I_n(\theta)$ par $I_O(\hat{\theta})$ évalué par différenciation finie. Ici, l'étape supplémentaire consiste à évaluer $\bigtriangledown g$. Plutôt que d'évaluer ce terme de manière analytique, car cela pourrait nous intéresser pour des fonctions g différentes, il vaut mieux évaluer cette dérivée en utilisant la différenciation finie; c'est-à-dire avec $i$ème élément de $\bigtriangledown g$ est approché par

$$\frac{\partial g(\theta)}{\partial \theta_i} \approx \frac{g(\hat{\theta} + \epsilon_i 1_i) - g(\hat{\theta})}{\epsilon_i}$$

pour un petit ϵ_i (comme cela s'applique quand $\epsilon_i \downarrow 0$). Le choix des paramètres de différenciation $\{\epsilon_i\}$ est important - vous devrez vérifier la cohérence sur une plage de valeurs pour ce vecteur. Souvent les $\{\epsilon_i\}$ sont sélectionnés par rapport à la précision de l'ordinateur ou sont automatiquement sélectionnés par le logiciel. Lorsque vous avez le contrôle sur eux, il est préférable de les considérer comme proportionnels aux estimations des paramètres $\epsilon_i / \hat{\theta}_i = c$. Cela protégera contre les erreurs relatives.

7. Simuler l'échantillon i.i.d de loi de Gamma(α, β) de taille $n = 200$ avec $\alpha = 2, \beta = 2$. Trouver l'estimateur et l'écart-type de l'estimateur de $\phi = \alpha/\beta$. Donner un intervalle de confiance à 95% de ϕ.

8. Simuler l'échantillons i.i.d de loi de Cauchy de paramètres $x_0 = 10$ and $\alpha = 0.1$. (rappel de la densité : $f(x; x_0, \alpha) = \frac{1}{\pi} \frac{\alpha}{(x - x_0)^2 + \alpha^2}$, avec $\alpha > 0$). Trouver l'estimateur et l'écart-type de l'estimateur pour: (i) $P(X > 100)$ et (ii) x tel que $P(X \leq x) = 0.99$. Donner un intervalle de confiance à 95% pour chaque cas.

Méthodes basées sur la simulation pour les intervalles de confiance

Pour les modèles complexes ou irréguliers, il n'est pas toujours possible d'obtenir la distribution thérorique ou asymptotique de l'estimateur.

Une autre façon de construire des intervalles de confiance pour l'estimateur consiste à utiliser la simulation. L'idée clé est que les intervalles de confiance sont déterminés à partir de la distribution des estimateurs obtenue à partir de réplications de l'échantillon initiale. En effet, si nous pouvons simuler de tels ensembles de données répliqués, alors pour chacun d'eux, nous pouvons calculer l'estimateur. Nous pouvons ensuite utiliser la distribution empirique résultante pour construire des intervalles de confiance appropriés.

L'un des principaux avantages de cette approche est que la distribution d'échantillonnage ne repose pas sur la théorie asymptotique. Cela peut être important si la taille des échantillons n'est pas suffisamment grande pour que les résultats asymptotiques soient précis, ou pour des modèles complexes où les résultats

asymptotiques standards ne sont pas valables. Le problème avec cette approche est que, comme nous ne connaissons pas la valeur réelle du paramètre, nous ne pouvons pas simuler des ensembles de données répliqués à partir de la distribution appropriée. Cependant, il existe deux approches pour simuler des ensembles de données répliqués à partir de distributions approximatives. Nous les appelons respectivement le bootstrap paramétrique et le bootstrap non paramétrique.

Bootstrap paramétrique

La distribution correcte pour simuler de nouveaux ensembles de données à partir de est $f(\mathbf{x}|\theta_0)$. L'idée du bootstrap paramétrique est que nous remplaçons θ_0 par son Estimateur du Maximum de Vraisemblance (EMV) $\hat{\theta}$. L'algorithme est alors le suivant:

(1) Simuler M ensembles de données (échantillons) indépendants , $\mathbf{x}_1, \ldots, \mathbf{x}_M$ à partir de $f(\mathbf{x}|\hat{\theta})$.
(2) Pour chaque échantillon $\mathbf{x}_i$, calculez l'EMV correspondant, $\hat{\theta}_i$.
(3) Construire un intervalle de confiance de 95% en utilisant les quantiles $2,5\%$ et $97,5\%$ de la distribution empirique de $\hat{\theta}_1, \ldots, \hat{\theta}_M$.

Bootstrap non-paramétrique

Le boostrap non paramétrique diffère uniquement dans la façon dont les ensembles de données répliqués sont simulés à l'étape (1). L'idée est de simuler de nouvelles données en rééchantillonnant, avec remise, les données réelles. Cela revient à simuler les échantillons à partir de la distribution empirique (que l'on peut visualiser par la fonction de répartition empirique qui donne un poids $1/n$ à chaque observation). Nous savons que cette approximation est bonne car la fonction de répartition empirique converge vers la vraie fonction de répartition empirique lorsque n croît.

Dans R, l'échantillonnage avec remise peut être mis en œuvre comme suit:

```
n = 10
xsim=sample(xdat,size=n,replace=T) ##sample with replacement
```

9. Soit $X_i \sim \mathrm{N}(\theta, 1)$ avec $\theta \geq 0$. Montrer que l'estimateur du maximum de vraisemblance est $\max(\bar{\mathbf{x}}, 0)$. Est-ce que c'est un modèle régulier? Construire des intervalles de confiance à 95% pour θ par (i) bootstrap paramétrique (ii) bootstrap non paramétrique. Comparer les résultats.

10. Pour l'échantillon i.i.d de loi de Gamma(α, β) de taille $n = 200$ avec $\alpha = 2, \beta = 2$, construire l' intervalle de confiance à 95% pour $\phi = \alpha/\beta$. Évaluer la couverture par (i) normalité asymptotique (ii) bootstrap paramétrique (iii) bootstrap non paramétrique.

11. Soit un mélange de lois normales défini par la densité

$$f(x|\mu, p) = pN(x|\mu) + (1-p)N(x|0)\,,$$

où $N(x|\mu)$ est la densité de $N(\mu, 1)$ évaluée en x et $0 < p < 1$. Simuler un échantillon i.i.d de taille $n = 100$. Construire des intervalles de confiance "bootstrap" à 95% pour μ et p.

Application aux données sur l'ozone

12. Pour les données sur l'ozone ("summer_ozone.csv", "winter_ozone.csv"), nous voulons savoir s'il y a une différence entre les deux sites à chaque saison. Soit θ_1 and θ_2 les différences moyennes du niveau d'ozone en été et en hiver respectivement. En utilisant les modèles que vous avez envisagés lors de la session précédente (ou autre si vous avez une proposition), construire des intervalles de confiance à 90% et 99% pour ces paramètres. Les intervalles incluent-ils zéro ?