

Quelques diagnostics locaux pour le modèle de Fay-Herriot

Éric Lesage

INSEE

24-26 octobre 2018

10^e Colloque francophone sur les sondages
Lyon

Travail réalisé en collaboration avec Cynthia Bocci et Jean-François Beaumont,
Statistique Canada

Contexte

- On considère une population U et un échantillon s tiré dans U selon un plan de sondage $p(s)$.
- U est partitionné en m domaines (domaines non planifiés) U_i ; $i = 1, \dots, m$.
- On s'intéresse aux paramètres d'intérêt θ_i associés aux domaines $i = 1, \dots, m$.
- On dispose d'information auxiliaire sous la forme de vecteurs \mathbf{z}_i de caractéristiques disponibles pour tous les domaines $i = 1, \dots, m$.

Modèle de liaison

- On suppose qu'on dispose d'un **modèle de liaison** qui nous permet de décomposer les paramètres d'intérêt θ_i de la façon suivante :

$$\theta_i = \boldsymbol{\beta}^\top \mathbf{z}_i + v_i,$$

où :

- (i) $\boldsymbol{\beta}^\top \mathbf{z}_i$ est l'effet connu ou expliqué par le modèle de θ_i ,
 - (ii) v_i est l'effet inconnu ou l'**effet local** de θ_i (qui ne dépend pas de \mathbf{z}_i).
- v_i est la réalisation d'une variable aléatoire qui suit une loi normale $\mathcal{N}(0, \sigma_v^2)$.
 - On fait l'hypothèse que $\boldsymbol{\beta}$ and σ_v^2 sont connus (pour la partie théorique de la présentation).

Estimateur direct pour le domaine i

- Il peut s'agir d'un estimateur Horvitz-Thompson ou un estimateur par calage.
- On note e_i l'erreur d'échantillonnage :

$$e_i = \hat{\theta}_i - \theta_i.$$

- On note ψ_i la variance de $\hat{\theta}_i$ sous le plan de sondage.
- Le nombre d'unités sélectionnées dans le domaine i , n_i , peut être très faible (voir nul)...
- ... ce qui peut conduire à une très faible précision de l'estimateur direct $\hat{\theta}_i$.

Estimateur composite pour le domaine i

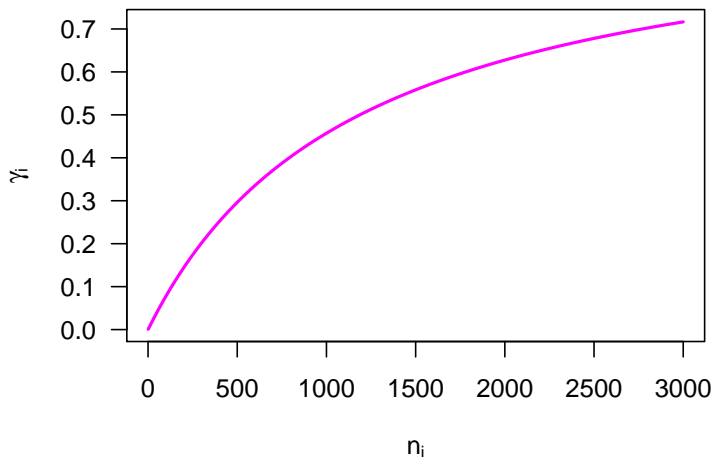
- Fay et Herriot ont proposé un **estimateur composite** dans le but d'obtenir un estimateur plus précis que l'estimateur direct :

$$\hat{\theta}_i^* = \gamma_i \hat{\theta}_i + (1 - \gamma_i) \beta^\top \mathbf{z}_i,$$

- $\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \tilde{\psi}_i}$ et $\tilde{\psi}_i = \mathbb{E}(\psi_i | \mathbf{z}_i)$.
- γ_i peut être vu comme une mesure de la force relative de l'échantillonnage par rapport au modèle de liaison.
 - $\gamma_i \rightarrow 1$, lorsque **la taille de l'échantillon est importante** où lorsque le modèle est très peu puissant.
 - $\gamma_i \rightarrow 0$, lorsque la taille de l'échantillon est faible où lorsque **le modèle est très puissant** (à la limite on pourrait se dispenser d'échantillonner le domaine)
 - $\gamma_i \approx 0,5$, lorsque l'erreur d'échantillonnage est du même ordre de grandeur que l'erreur du modèle de liaison.

Illustration du lien entre γ_i et n_i pour un domaine i donné

Relation between γ_i and n_i



Inférence sous le modèle combiné ou sous le plan

- L'inférence sous le plan de sondage et le modèle de liaison est appelée inférence sous le **modèle combiné**. On peut écrire ce modèle sous la forme :

$$\hat{\theta}_i = \beta^\top \mathbf{z}_i + v_i + e_i.$$

- Afin d'apprécier la précision de l'estimateur composite, on a deux possibilités :
 - soit prendre l'**erreur quadratique moyenne sous le plan de sondage**,
 - soit prendre l'**erreur quadratique moyenne sous le modèle combiné**.
- L'approche sous le modèle combiné, fournit les EQM suivantes pour les estimateurs direct et indirect du domaine i :
 - (i) $mse(\hat{\theta}_i | \mathbf{z}_i) = \tilde{\psi}_i$
 - (ii) $mse(\hat{\theta}_i^* | \mathbf{z}_i) = \gamma_i \tilde{\psi}_i$
- **L'estimateur composite est donc toujours plus précis que l'estimateur direct avec une inférence sous le modèle combiné** (ce résultat résulte de la construction même de l'estimateur composite)

Inférence sous le plan

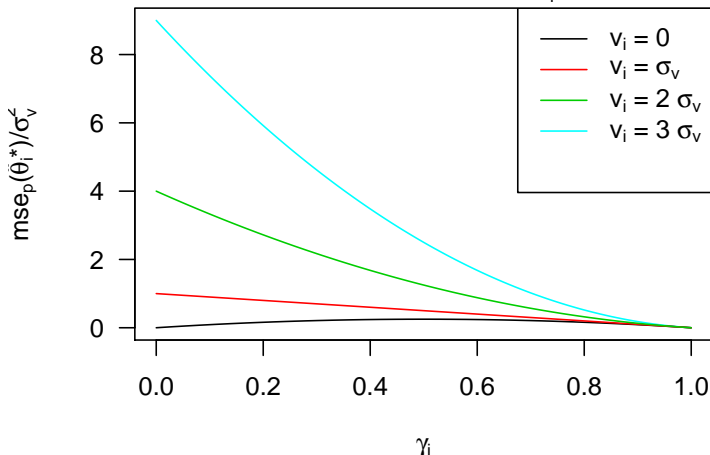
- Par contre, et cela est une question légitime, on peut se demander si l'estimateur composite est toujours plus précis que l'estimateur direct avec une **inférence sous le plan de sondage** ?
- On montre que ça dépend de l'amplitude de l'effet local v_i et éventuellement de la valeur du coefficient γ_i .
- On donnera **deux diagnostics locaux** conçus pour évaluer si l'estimateur composite est susceptible d'être plus précis que l'estimateur direct. On entend par diagnostic local un diagnostic qui porte sur un domaine spécifique.

EQM sous le plan des estimateurs direct et indirect

- Les EQM sous le plan des estimateurs direct et indirect pour le domaine i valent :
 - (i) $mse_p(\hat{\theta}_i) = \psi_i$
 - (ii) $mse_p(\hat{\theta}_i^*) = \gamma_i^2 \psi_i + (1 - \gamma_i)^2 v_i^2$
- **Pour un domaine i donné et une valeur v_i fixée**, on analyse les valeurs de $mse_p(\hat{\theta}_i)$ et $mse_p(\hat{\theta}_i^*)$ en fonction des valeurs de γ_i , Sous l'hypothèse simplificatrice $\psi_i \approx \tilde{\psi}_i$ on obtient :
 - (i) $mse_p(\hat{\theta}_i) \approx \frac{1 - \gamma_i}{\gamma_i} \sigma_v^2$,
 - (ii) $mse_p(\hat{\theta}_i^*) \approx (1 - \gamma_i) \sigma_v^2 + (1 - \gamma_i)^2 (v_i^2 - \sigma_v^2)$ ou
 $mse_p(\hat{\theta}_i^*) \approx (1 - \gamma_i) \gamma_i \sigma_v^2 + (1 - \gamma_i)^2 v_i^2$
- On en déduit que :
 - (i) lorsque $|v_i| < \sqrt{2} \sigma_v$, $mse_p(\hat{\theta}_i^*)$ est plus petit que $mse_p(\hat{\theta}_i) \forall \gamma_i$.
 - (ii) A l'inverse, lorsque $|v_i| > \sqrt{2} \sigma_v$, il existe une valeur de γ_i au delà de laquelle $mse_p(\hat{\theta}_i^*) > mse_p(\hat{\theta}_i)$.

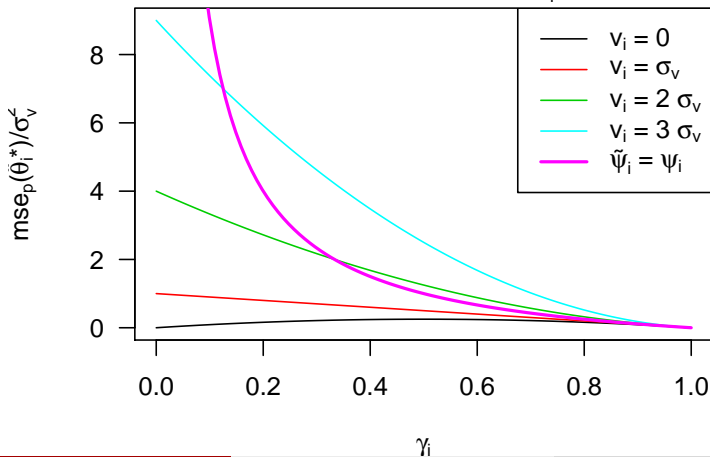
$mse_p(\hat{\theta}_i^*)$ peut être supérieure à $mse_p(\hat{\theta}_i)$

Design MSE of the Composite Estimator: $mse_p(\hat{\theta}_i^*)$
for different values of v_i



$mse_p(\hat{\theta}_i^*)$ peut être supérieure à $mse_p(\hat{\theta}_i)$

Design MSE of the Composite Estimator: $mse_p(\hat{\theta}_i^*)$
for different values of v_i



L'effet local v_i doit donc être d'ampleur limitée si on espère obtenir que $mse_p(\hat{\theta}_i^*) \leq mse_p(\hat{\theta}_i)$

On raisonne maintenant à γ_i fixé pour un domaine i donné.

Definition (Limites de v_i)

$v_i(-)$ et $v_i(+)$ sont les deux valeurs limites telles que :

$$mse_p(\hat{\theta}_i^*) \leq mse_p(\hat{\theta}_i) \iff v_i \in [v_i(-), v_i(+)]$$

Proposition (Limites de v_i)

- (i) $v_i(+)$ et $v_i(-)$ $\approx \sigma_v \sqrt{\frac{1 + \gamma_i}{\gamma_i}}$, sous l'hypothèse $\psi_i \approx \tilde{\psi}_i$.
- (ii) Pour des grandes valeurs de γ_i (i.e. γ_i proche de 1), le paramètre $|v_i|$ doit être inférieur à $\sqrt{2} \sigma_v$ pour avoir $mse_p(\hat{\theta}_i^*) \leq mse_p(\hat{\theta}_i)$.
- (iii) Pour de petites valeurs de γ_i , le paramètre $v_i(+)$ peut prendre de très grandes valeurs !

Pour un domaine i donné

Limit values for the local effect v_i

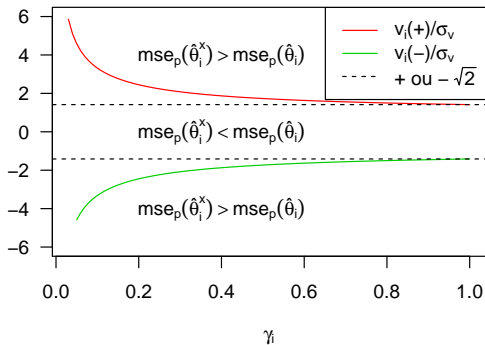


Figure: Valeurs limites de l'effet local en fonction de γ_i

Deux diagnostics pour choisir entre l'estimateur direct et l'estimateur indirect

- Pour un domaine i donné, le choix entre l'estimateur direct et l'estimateur indirect dépend de :
 - (i) La valeur du coefficient γ_i qui est connue (ou estimée) ;
 - (ii) La valeur de v_i , qui est certes **inconnue** mais qui est captée indirectement à travers la quantité observée : $\hat{\theta}_i - \beta^\top \mathbf{z}_i = e_i + v_i$.
- On utilise deux approches différentes pour dériver les deux diagnostics destinés à décider quel estimateur est le plus précis sous le plan de sondage :
 - (i) La première approche s'appuie sur des **tests statistiques d'hypothèses** qui concernent le paramètre v_i dans le cadre d'une inférence sous le plan.
 - (ii) La seconde approche dite "**Empirical Bayes**" (EB) s'appuie sur une distribution conditionnelle de v_i .

Première approche : test d'hypothèses sous le plan

- v_i est considéré comme un **paramètre** inconnu.
- Sous le plan de sondage, la variable aléatoire $\hat{\theta}_i - \beta^\top \mathbf{z}_i$ suit la loi :

$$\hat{\theta}_i - \beta^\top \mathbf{z}_i \mid v_i, \mathbf{z}_i \sim \mathcal{N}(v_i, \psi_i),$$

- Nous avons une unique observation de cette variable aléatoire $\hat{\theta}_i - \beta^\top \mathbf{z}_i$;
- On utilise deux tests pour évaluer si $|v_i|$ est plus petit que $v_i(+)$

(i) **Deux jeux d'hypothèses :**

$$\begin{array}{ll} \mathbf{H}_0(+): v_i = v_i(+) & \mathbf{H}_1(+): v_i > v_i(+) \\ \mathbf{H}_0(-): v_i = v_i(-) & \mathbf{H}_1(-): v_i < v_i(-) \end{array}$$

(ii) **Statistiques de test:**

$$S_i(+) = \frac{\varepsilon_i - \sqrt{1 + \gamma_i}}{\sqrt{1 - \gamma_i}} \times \sqrt{\frac{\tilde{\psi}_i}{\psi_i}} \sim \mathcal{N}(0, 1)$$

$$S_i(-) = \frac{\varepsilon_i + \sqrt{1 + \gamma_i}}{\sqrt{1 - \gamma_i}} \times \sqrt{\frac{\tilde{\psi}_i}{\psi_i}} \sim \mathcal{N}(0, 1)$$

où $\varepsilon_i = \frac{\hat{\theta}_i - \beta^\top \mathbf{z}_i}{\sqrt{\sigma_v^2 + \tilde{\psi}_i}}$ est l'erreur standardisée du domaine i .

Première approche : test d'hypothèses sous le plan

(iii) P-values:

$$Pvalue(+) = \Phi(-S_i(+))$$

$$Pvalue(-) = \Phi(S_i(-))$$

où $\Phi(\cdot)$ est la fonction de répartition de la loi normale.

Definition (Premier diagnostic)

Le premier diagnostic est défini comme le minimum des deux P-values ($Pvalue(-)$, $Pvalue(+)$) :

$$Diag(D)_i = \Phi\left(\frac{\sqrt{1+\gamma_i} - |\varepsilon_i|}{\sqrt{1-\gamma_i}}\right)$$

où $\varepsilon_i = \frac{\hat{\theta}_i - \beta^T \mathbf{z}_i}{\sqrt{\sigma_v^2 + \psi_i}}$ est l'erreur standardisée du domaine i .

Première approche : test d'hypothèses sous le plan

- **Le premier diagnostic s'interprète de la façon suivante :**
 - (i) Quand $Diag(D)_i$ est faible, on rejète H_0 , ce qui signifie qu'on "garde" l'estimateur direct $\hat{\theta}_i$.
 - (ii) A l'inverse, lorsque $Diag(D)_i$ est grand, on accepte H_0 , ce qui signifie qu'on "garde" l'estimateur composite $\hat{\theta}_i^*$.
- **Pour obtenir une règle de décision il est nécessaire de choisir un seuil :**
 - (i) On peut par exemple s'inspirer des valeurs utilisées habituellement comme niveaux pour les tests ($\alpha = 5\%$ ou 10%). Avec ces valeurs, compte tenu du sens du test, on favorise l'estimateur composite.
 - (ii) Une autre idée est d'adopter une approche empirique et de regarder la distribution des valeurs du diagnostic pour les m domaines $\{Diag(D)_i, i \in (1, \dots, m)\}$ et de repérer une rupture.

Seconde approche : approche Empirical Bayes

- Dans cette approche, v_i est vue comme la réalisation d'une variable aléatoire qui suit une loi normale $\mathcal{N}(0, \sigma_v^2)$.
- Selon l'approche "Empirical Bayes" (Rao et Molina, 2015, chap 9, pages 271-272), nous avons la loi conditionnelle de v_i :

$$v_i \mid \hat{\theta}_i - \beta^\top \mathbf{z}_i \sim \mathcal{N} \left(\gamma_i (\hat{\theta}_i - \beta^\top \mathbf{z}_i), (1 - \gamma_i) \sigma_v^2 \right).$$

- Le second diagnostic est défini comme la probabilité conditionnelle d'avoir $mse_p(\hat{\theta}_i^*) \leq mse_p(\hat{\theta}_i)$, c'est-à-dire:

$$\begin{aligned} \text{Diag}(EB)_i &= \text{Prob} \left(mse_p(\hat{\theta}_i^*) \leq mse_p(\hat{\theta}_i) \mid \hat{\theta}_i - \beta^\top \mathbf{z}_i \right) \\ &= \text{Prob} \left(v_i(-) \leq v_i \leq v_i(+) \mid \hat{\theta}_i - \beta^\top \mathbf{z}_i \right) \\ &= \Phi \left(\frac{v_i(+)-\gamma_i(\hat{\theta}_i-\beta^\top \mathbf{z}_i)}{\sqrt{1-\gamma_i} \sigma_v} \right) - \Phi \left(\frac{v_i(-)-\gamma_i(\hat{\theta}_i-\beta^\top \mathbf{z}_i)}{\sqrt{1-\gamma_i} \sigma_v} \right) \end{aligned}$$

Seconde approche : approche Empirical Bayes

Definition (Second diagnostic)

$$\begin{aligned} \text{Diag}(EB)_i &= \Phi \left\{ \sqrt{\frac{\gamma_i}{1-\gamma_i}} \left(|\varepsilon_i| + \frac{\sqrt{1+\gamma_i}}{\gamma_i} \right) \right\} - 1 \\ &+ \Phi \left\{ \sqrt{\frac{\gamma_i}{1-\gamma_i}} \left(\frac{\sqrt{1+\gamma_i}}{\gamma_i} - |\varepsilon_i| \right) \right\}, \end{aligned}$$

où $\varepsilon_i = \frac{\hat{\theta}_i - \beta^\top \mathbf{z}_i}{\sqrt{\sigma_v^2 + \tilde{\psi}_i}}$ est l'erreur standardisée du domaine i .

Etude par simulation

- On réalise l'étude par simulation à partir d'un jeu de données réelles de *l'enquête emploi canadienne de mai 2011* :
 - (i) $m = 140$ aires métropolitaines de recensement et aires de recensement
 - (ii) z_i est le ratio du nombre d'allocataires de l'assurance chômage dans l'aire i sur le nombre de personnes de plus de 15 dans l'aire i .
 - (iii) n_i est la taille de l'échantillon dans l'aire i

Etude par simulation

- Ensuite, à partir de ces données réelles, on simule les paramètres θ_i pour les m domaines.
 - Les paramètres d'intérêt θ_i peuvent être interprétés comme des taux de chômage des m domaines.
 - On réalise trois jeux de simulations pour les θ_i : un jeu qu'on appellera jeu initial, puis un second jeu appelé "premier jeu modifié" et enfin un troisième jeu qualifié de "second jeu modifié".
 - Dans le **jeu initial**, θ_i suit une loi bêta $Beta(\beta^T \mathbf{z}_i, \sigma_v^2)$, où $\sigma_v = 0.87\%$ et $\beta^T = (0.0484, 1.95)$.
 - Dans le **premier jeu modifié** ($5\sigma_v$) : on a modifié manuellement les valeurs de θ_i pour 4 aires de manière à avoir $v_i = 5 \times \sigma_v$.
 - Dans le **second jeu modifié** ($15\sigma_v$) : on a modifié manuellement les valeurs de θ_i des mêmes 4 aires de manière à avoir $v_i = 15 \times \sigma_v$.
- Le plan de sondage est un tirage aléatoire simple stratifié par aire avec remise. ψ_i and $\tilde{\psi}_i$ sont calculés une fois unique pour chaque $i = 1, \dots, m$.

Simulations Monte Carlo Simulation avec $K = 10\,000$ replications

- On répète $K = 10\,000$ fois l'échantillonnage.
- Pour chaque itération k , $k = 1, \dots, K$, on sélectionne un échantillon et on calcule :
 - L'estimateur direct $\hat{\theta}_i(k)$
 - l'estimateur composite $\hat{\theta}_i^*(k)$:

$$\hat{\theta}_i^*(k) = \hat{\gamma}_i(k) \hat{\theta}_i(k) + \{1 - \hat{\gamma}_i(k)\} \hat{\beta}(k)^\top \mathbf{z}_i$$

$$(i) \hat{\gamma}_i(k) = \frac{\hat{\sigma}_v^2(k)}{\hat{\sigma}_v^2(k) + \tilde{\psi}_i}$$

- (ii) $\hat{\beta}(k)$ et $\hat{\sigma}_v(k)$ sont estimés (moindres carrés pondérés pour $\hat{\beta}$ et maximum de vraisemblance réduite (REML) pour $\hat{\sigma}_v$).

- On calcule également les quantités Monte Carlo $\hat{v}_i(+)(k)$, $\hat{v}_i(-)(k)$, $\hat{\varepsilon}_i(k)$, $\widehat{Diag}(EB)_i(k)$ et $\widehat{Diag}(D)_i(k)$.

Mesures Monte Carlo

- $\overline{Diag(EB)}_i$ et $\overline{Diag(D)}_i$ sont les moyennes Monte Carlo des diagnostics calculés à l'occasion des $K = 10\ 000$ itérations.
- La mesure Monte Carlo de l'erreur quadratique sous le plan de sondage de l'estimateur $\hat{\theta}_i^*$ vaut :

$$mse_{MC}(\hat{\theta}_i^*) = \frac{1}{K} \sum_{k=1}^K (\hat{\theta}_i^*(k) - \theta_i)^2.$$

- On calcule la différence relative entre la mesure Monte Carlo de l'EQM sous le plan de l'estimateur composite et l'EQM sous le plan de l'estimateur direct $\hat{\theta}_i$:

$$\frac{mse_{MC}(\hat{\theta}_i^*) - \psi_i}{\psi_i}.$$

- Quand ce ratio est positif, on a un estimateur indirect moins précis que l'estimateur direct. On va comparer la valeur moyenne des deux diagnostics à ce ratio. Le diagnostic sera donc performant si les valeurs qu'il prend pour chacun des domaines sont corrélées aux valeurs du ratio pour ces mêmes domaines.

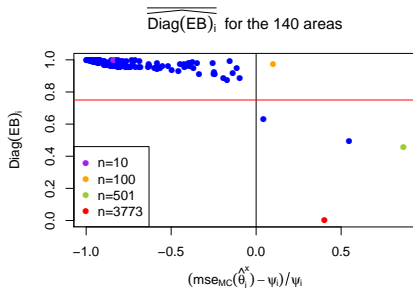
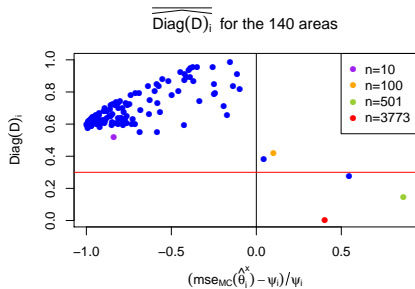


Figure: diagnostics D et EB Monte Carlo avec le premier jeu modifié ($5\sigma_v$)

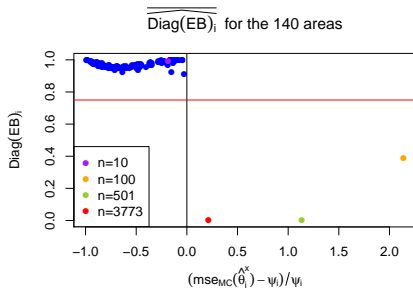
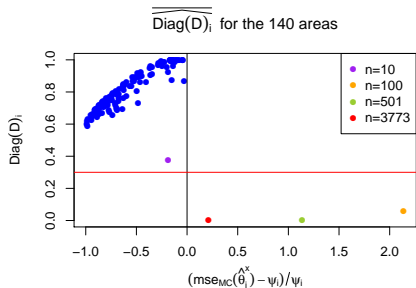


Figure: diagnostics D et EB Monte Carlo avec le second jeu modifié ($15 \sigma_v$)

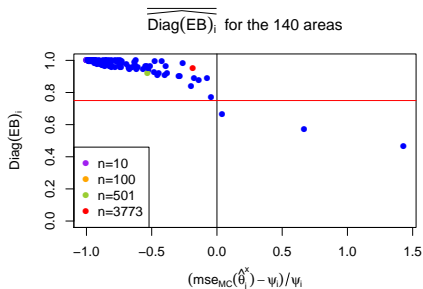
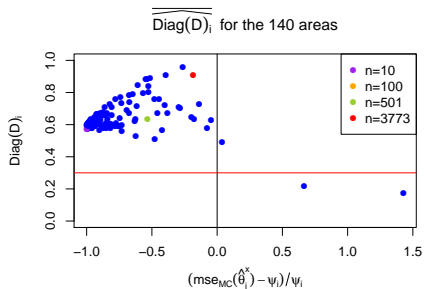


Figure: diagnostics D Monte Carlo avec le jeu initial