Statistique et aide à la décision Session 4

David Causeur Agrocampus Ouest IRMAR CNRS UMR 6625

Plan

1 Comparaison de modèles

Choix du meilleur modèle

Test du rapport de vraisemblance

Illustration. Le lien entre l'indice de couleur *a* et la maturité est-il le même pour toutes les variétés d'abricots ?

Modèle \mathcal{M}_1 pour la maturité d'un abricot :

$$\begin{cases} \log \frac{\pi_{2i}(x)}{\pi_{1i}(x)} &= \mu^{(2)} + \alpha_i^{(2)} + (\beta^{(2)} + \gamma_i^{(2)})X \\ \log \frac{\pi_{3i}(x)}{\pi_{1i}(x)} &= \mu^{(3)} + \alpha_i^{(3)} + (\beta^{(3)} + \gamma_i^{(3)})X \end{cases}$$

où $\pi_{ki}(x)$ est la probabilité qu'un abricot de variété i, dont l'indice a vaut x, soit dans le stade k de maturité.

Sous H_0 , $\gamma_i^{(k)} = 0$, pour tout i, pour tout k.



Test du rapport de vraisemblance

Comparaison des modèles \mathcal{M}_0 et \mathcal{M}_1 :

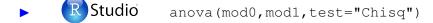
 \mathcal{M}_0 : sous-modèle de \mathcal{M}_1 obtenu avec $\gamma_i = 0$, pour tout i.

Statistique de test du rapport de vraisemblance :

LRT = $\mathcal{D}_0 - \mathcal{D}_1$, où \mathcal{D}_0 est la déviance résiduelle de \mathcal{M}_0 et \mathcal{D}_1 est la déviance résiduelle de \mathcal{M}_1 .

Loi de LRT sous H₀:

LRT $\sim_{_{\mathrm{H}_{\mathrm{0}}}} \chi_{q}^{2}$, où q est la différence entre les nombres de paramètres de \mathcal{M}_{1} et \mathcal{M}_{0} .



Plan

Comparaison de modèles

2 Choix du meilleur modèle

Quel modèle \mathcal{M}_k avec $k \leq K$ variables explicatives est suffisant pour expliquer le stade de maturité ?

Choix parmi 2^K modèles non-emboîtés

Le Critère d'Information d'Akaike $AlC_{x,y}(\hat{\beta})$ est défini ainsi :

$$AIC_{X,V}(\hat{\beta}) = \mathcal{D}_{X,V}(\hat{\beta}) + 2p,$$

où p est le nombre de paramètres du modèle.

 $AIC_{x,y}(\hat{\beta})$ estime la perte d'information lorsque l'on utilise le modèle estimé avec $\hat{\beta}$ plutôt que le vrai modèle ayant généré les données.

Le Critère d'Information Bayésien $BIC_{x,y}(\hat{\beta})$ est défini ainsi :

$$BIC_{x,y}(\hat{\beta}) = \mathcal{D}_{x,y}(\hat{\beta}) + p \ln(n),$$

où p est le nombre de paramètres du modèle.

 $BIC_{x,y}(\hat{\beta})$ estime la perte d'information lorsque l'on utilise le modèle estimé avec $\hat{\beta}$ plutôt que le vrai modèle ayant généré les données dans le périmètre des modèles paramétriques considérés.

Quel modèle \mathcal{M}_k avec $k \leq K$ variables explicatives est suffisant pour expliquer le stade de maturité?

Choix parmi 2^K modèles non-emboîtés

- Parmi les modèles M_k: le modèle M_k* avec la plus petite déviance résiduelle D_k* est le champion
- Le champion des champions M* est le modèle M_k* avec le plus petit AIC (ou BIC)
- bestglm(Xy,family,method,IC)
- ► R Studio

AIC ou BIC?

- Si le but est de prédire, minimiser AIC est recommandé.
- Si le but est d'ajuster un modèle aux données, minimiser BIC doit être privilégié.

Minimiser BIC conduit à choisir des modèles plus parcimonieux

Algorithmes sous-optimaux

Si K > 15 explanatory variables, recherche forward (ou backward) stepwise

- Etape 1 : M₁*
- Etape $k: \mathcal{M}_k^{\star}$ parmi les modèles complétant $\mathcal{M}_{k-1}^{\star}$ en ajoutant une variable explicative.
- Stop si le BIC de \mathcal{M}_k^{\star} est plus grand que celui de $\mathcal{M}_{k-1}^{\star}$.
- stepwise(mod, direction, criterion)
- Studio