Statistique et aide à la décision
Session 4

David Causeur
Agrocampus Ouest
IRMAR CNRS UMR 6625

Plan

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.

- RStudio

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 H_0 :

LRT $\sim_{H_0}$ χ^2_q, où q est la différence entre les nombres de paramètres de $\mathcal{M}_1$ et $\mathcal{M}_0$.

► RStudio `anova(mod0,mod1,test="Chisq")`

Plan

Sélection de variables explicatives

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

Sélection de variables explicatives

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

$$\mathrm{AIC}_{x,y}(\hat{\beta}) = \mathcal{D}_{x,y}(\hat{\beta}) + 2p,$$

où p est le nombre de paramètres du modèle.

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

Sélection de variables explicatives

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

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

où p est le nombre de paramètres du modèle.

$\mathrm{BIC}_{x,y}(\hat{\beta})$ 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.

Sélection de variables explicatives

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 $\mathcal{M}_k$: le modèle $\mathcal{M}_k^\star$ avec la plus petite déviance résiduelle $\mathcal{D}_k^\star$ est le champion

- Le champion des champions $\mathcal{M}^\star$ est le modèle $\mathcal{M}_k^\star$ avec le plus petit AIC (ou BIC)

- `bestglm(Xy,family,method,IC)`
- R Studio

Sélection de variables explicatives

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** : $\mathcal{M}_1^\star$
- **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)`
- ► RStudio