Scénario : modélisation de nids de processionnaires du pin

Résumé

Modélisation par régression linéaire simple puis multiple avec sélection de variables du nombre de nids de processionnaires du pin. C'est un cas de forte colinéarité introduisant aux procédures de sélection de variables dans le modèle gaussien. Les données sont traitées avec SAS.

1 Introduction

1.1 Problématique

La processionnaire du pin (Thaumetopea pityocampa) est un insecte de l'ordre des lépidoptères. Les larves sont connues pour leur mode de déplacement en file indienne, se nourrissent des aiguilles de diverses espèces de pins, provoquant un affaiblissement important des arbres. Si leurs longs poils (soies) sont inoffensifs, ces chenilles projettent dans l'air de minuscules poils très urticants à partir du 3ème stade larvaire. Leur fort caractère urticant peut provoquer d'importantes réactions allergiques (cf. Wikipédia).

L'étude proposée s'inspire d'un exemple extrait du livre de Tomassone et col. (1992) [1]. L'objectif est d'étudier l'influence de certaines caractéristiques de peuplements forestiers sur la densité de la processionnaire et, plus précisément, on souhaite construire un modèle prédictif d'une variable mesurant cette densité afin de fournir des recommandations aux forestiers dans la conduite et la surveillance des zones forestières.

1.2 Les données

L'unité, qui représente ici l'observation, est une parcelle forestière de 10 hectares d'un seul tenant. On a une seule valeur de chaque variable pour chaque parcelle. En fait, pour tenir compte d'un éventuel manque d'homogénéité, dans chaque parcelle, on a mesuré les variables sur plusieurs placettes de 5 ares chacune. Et la valeur attribuée à la parcelle est la moyenne des valeurs obtenue pour ses placettes.

Les variables observées pour chaque parcelle sont les suivantes :

`Alti` altitude (en m)

`Pent` pente (en degrés)

`NbPi` nombre de pins par placette

`Haut` hauteur (en m) du pin central de la placette

`Diam` diamètre de ce pin

`Dens` note de densité de la végétation de la placette

`Orie` orientation (de 1 vers le sud à 2 vers le nord)

`Hdom` hauteur (en m) de l'arbre dominant

`Stra` nombre moyen de strates de végétation

`Mela` mélange du peuplement (de 1, pas mélangé à 2, mélangé)

`NbNi` densité de processionnaire (nombre moyen de nids par arbre)

Le travail va consister à répondre à une liste de questions en s'aidant des sorties de différentes commandes SAS.

2 Étude préalable

Quelle est la population étudiée ? Quelle est la taille de l'échantillon ?

2.1 Lecture des données

Prise en charge des données : charger les données du fichier `proc-pin.dat` à partir de l'URL : `http://wikistat/data/`

Exécuter le programme de lecture :

```
data sasuser.procpin;
infile "proc-pin.dat";
input Alti Pent NbPi Haut Diam Dens Orie Hdom Stra
      Mela NbNi ;
run;
```

1. *La régression : nouveaux regards sur une ancienne méthode*, 2ème édition, Masson

2.2 Exploration des données

```
solutions > Analysis > Interactive Data analysis
sasuser
procpin > open
```

Que dire sur la distribution de la variable NbNi ?

```
Analyse > Distribution(Y) > LNbNi > Y > OK
```

Contrôler rapidement les distributions des autres variables (symmétrie, valeurs atypiques).

Tester la normalité de NbNi avec la fenêtre NbNi active :

```
Tables > Test for normality
```

Calculer les transformations de NbNi par les fonctions log et racine :

```
NbNi > Edit > Variables > Log(Y)
NbNi > Edit > Variables > Sqrt(Y)
```

Etudier les distributions de ces nouvelles variables, quelle transformation vous semble la plus pertinente ?

Calculer la matrice des corrélations :

```
Analyse > Multivariate > tout sélectionner > Y > OK
```

Tracer la matrice des nuages de points (*scatter plot matrix*) :

```
Analyse > Scatter plot > tout sélec. >
   Y tout sélec. > X > OK
```

Que dire des natures des liaisons, quelles variables semblent les plus appropriées pour expliquer le nombre de nids.

Calculer et représenter l'analyse en composantes principales :

```
Analyse>Multivariate>pas sélec NnNi et S_NbNi>Y
Output>Principal Component Options
Principal component options >
   Biplot (Std Y)>OK>OK>OK
```

Commenter la structure de corrélation des variables.

3 Modèles de régression

L'étude est présentée selon une progression "pédagogique" à complexité croissante : régression simple, régression multiple, sélection de variables. Ce n'est pas la façon "nomimale" de conduire une analyse statistique qui nécessiterait d'étudier tout de suite le modèle complet (régression multiple) correspondant précisément à l'expérimentation réalisée (*experimental design*).

3.1 Régression simple

Il a été choisi d'expliquer le nombre de nids par l'altitude. Justifier ce choix.

Estimer successivement chacun des modèles expliquant NbNi, S_NbNi, L_NbNi par la variable Alti.

```
Analyse>fit, NbNi dans Y, Alti dans X
Output > Collinearity Diagnostics >
    Residual Normal QQ
Output variables>Studentized residuals>
    Cook's D >OK>OK>OK
```

Vérifier pour chacun d'eux les hypothèses de
- linéarité (graphe des résidus)
- homoscédasticité (graphe des résidus)
- normalité des résidus (droite de Henri, test de la variable)
- contrôler l'existence ou non de points influents (graphe distance de Cook)

Que dire de la qualité d'ajustement des modèles ? Quel transformation ou modèle vous semble le plus raisonnable ?

Sous réserve de validité de ces hypothèses, conclure sur l'influence de l'Altitude sur le nombre de nids.

3.2 Modèle complet

Analyse interactive

Estimer(Analyse>fit les modèles complets expliquant NbNi, S_Nbni, L_NbNi (sélectionnée successivement en Y) et avec toutes les variables sélectionnées en X.

Vérifier les hypothèses et diagnostics (linéarité, homoscédasticité, normalité des résidus, points influents) comme pour la régression simple.

Quel modèle choisir (regarder les tests de normalité des résidus) ? Que dire de la qualité d'ajustement ? Comparer à celle de la régression simple. Etudier les p-valeurs des différents tests, que conclure ?

Programme SAS

Exécuter les mêmes analyses sur plusieurs jeux de données nécessite d'automatiser les traitements. Retrouver les mêmes résultats fournis par la procédure classique SAS/REG dans le programme suivant. Beaucoup d'options y sont actives afin de fournir exhaustivement tous les résultats même si certains sont redondants ou peu utiles. Faire attention aux résidus "studentisés" qui prennent différentes formes de normalisation.

```
data procpin;
set sasuser.procpin;
SNbNi=sqrt(NbNi);
run;
proc reg data=procpin;
model SNbNi = Alti--Mela
    /dw covb Influence cli clm tol vif collin R P;
run;
```

Que sont les VIF (*variance influence factors*) Que dire de l'indicateur `collin` ? Quelle leçon tirer de ces résultats ?

4 Sélection de modèle

4.1 Choix de modèle "à la main" par élimination

SAS propose des algorithmes de sélection automatique des variables. Néanmoins il est nécessaire de savoir se "débrouiller" avec les outils plus limités proposés par d'autres logiciels.

Après avoir estimé comme précédemment le modèle complet, itérer la procédure suivante dans SAS/INSIGHT :

1. Choisir, parmi les variables explicatives, celle X^j pour lequel le test de Student ($H_0 : b_j = 0$) est le moins significatif, c'est-à-dire avec la plus grande "prob value".
2. La retirer du modèle et recalculer l'estimation. Il suffit pour cela de sélectionner le nom de la variable dans le tableau (TYPE III) et d'exécuter la commande `delete` du menu `edit` de la même fenêtre. Le modèle est ré-estimé automatiquement.

Arrêter le processus lorsque tous les coefficients sont considérés comme significativement (à 10%) différents de 0. Attention, la "variable" INTERCEPT (terme constant) ne peut pas être considérée au même titre que les autres variables ; la *conserver* toujours dans le modèle.

Noter le modèle finalement obtenu, son coefficient de détermination à comparer à celui du modèle complet.

4.2 Procédures automatiques

Comparer avec les modèles obtenus par différentes procédures de sélection (descendante, ascendante, pas à pas). Estimer chaque modèle séparément.

```
proc reg data=procpin;
model LNbNi = Alti--Mela / selection=backward;
run;
proc reg data=procpin;
model SNbNi = Alti--Mela / selection=forward;
run;
proc reg data=procpin;
model SNbNi = Alti--Mela / selection=stepwise;
run;
```

4.3 Optimisation globale

les algorithmes précédents peuvent, ou pas, passer par le "meilleur" modèle. Parmi les trois types d'algorithmes disponibles dans SAS et les différents critères de choix, une des façons les plus efficaces consistent à choisir les options du programme ci-dessous. Tous les modèles (parmi les plus intéressants selon l'algorithme de Furnival et Wilson) sont considérés. Seul le meilleur pour chaque niveau, c'est-à-dire pour chaque valeur q du nombre de variables explicatives sont donnés. Il est alors facile de choisir celui minimisant l'un des critères globaux (C_p, BIC...) estimant un risque pénalisé. Cette procédure de recherche d'un modèle optimal global n'est exécutable que pour un nombre

raisonnables de variables, disons moins d'une vingtaine.

```
proc reg data=procpin;
model SNbNi = Alti--Mela
   / selection=rsquare cp adjrsq bic best=1;
run;
```

Sélectionner le modèle de C_p minimum et celui de R^2 ajusté maximum. Comparer les sélections obtenues.

4.4 Comparaisons

Parmi les sélections précédentes, déterminer quelles sont celles différentes : choix descendant, descendant, stepwise, C_p minimum, R^2 ajusté maximum. Recalculer chacun de ces modèles :

```
proc reg data=procpin;
model LNbNi =   Alti Pent Haut Diam Dens Stra
     / collin r p ;
run;
proc reg data=procpin;
model LNbNi =   Alti Pent Haut Diam Dens Orie Stra
     / collin r p ;
run;
```

Comparer dans un tableau les valeurs obtenues des coefficients de détermination et de PRESS. Quel modèle choisir pour la prévision ?