Master Stat-Eco 2e année. Régression

Examen du 13 décembre 2011. Durée 2 heures.
Ni téléphone ni calculatrice.

Exercice 1 Un goûteur teste des chocolats fabriqués à base de cacao de trois provenances différentes : Côte d'Ivoire, Venezuela, Brésil. Il donne une note pour chaque chocolat qu'il goûte. Les chocolats sont préparés avec des doses de vanilline différentes. Proposer pour cette expérience un modèle de régression avec interaction et un sans interaction. Combien ont-ils de paramètres ? Interpréter leur différence.

Solution. Fait en TD. L'énoncé indiquant que les doses sont différentes, la dose est une variable quantitative. Ceci fait respectivement 4 paramètres (une moyenne et 3 pentes) et 6 paramètres (3 moyennes et 3 pentes), variance non comprise. Le deuxième modèle autorise que l'infuence de la vanille sur le goût diffère selon la provenance du chocolat.

Exercice 2 On drogue 48 rats avec des poisons de nocivité variable. Quatre traitements sont indifféremment utilisés pour les soigner. Les variables sont :

- y : inverse de la durée de survie du rat
- Poison : nocivité du poison
- Traitement (quatre modalités)

1. La commande `anova(lm(y∼Poison*Traitement))` donne les résultats suivants :

	Dl	Sum Sq	Pr(>F)
Poison	1	24,5	1,2e-8
Traitement	3	20,4	1,9e-06
Poison :Traitement	3	1,5	0,38
Residuals	40	19,1	

   Quelles conclusions cette table suggère-t-elle ? (En français clair)

   Solution. Le poison a un effet significatif sur la durée de vie, et les traitements ont des effets différents. Cette différence entre traitements est similaire quelle que soit la nocivité du poison.

2. La commande `summary(lm(y∼Poison+Traitement))` donne les résultats suivants :

	Estimate	Std.	Pr(>|t|)
(Intercept)	5,01	0,29	< 2e-16
Traitement2	-1,51	0,28	5,8e-07
Traitement3	-0,27	0,28	0,09
Traitement4	-1,38	0,28	1,9e-05
Poison	-0,25	0,05	8,1e-09

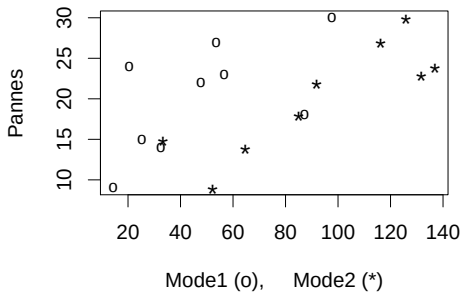
   Quelles conclusions cette table suggère-t-elle ?

   Solution. Les traitements 1 et 3 ont un effet similaire. De même pour les traitements 2 et 4, puisque les intervalles de confiance semblent se chevaucher largement. Ceci peut être confirmé par un test de ce modèle contre un modèle ou l'on fusionne séparément les deux paires de traitements.

3. Discuter de l'opportunité du choix du type de modèle utilisé..

   Solution. La durée de vie étant une variable positive, il serait bon d'essayer soit un changement de variable logarithmique, soit un modèle linéaire généralisé avec loi gamma ou inverse gaussienne.

Exercice 3 On dispose des données suivantes sur les pannes d'une pièce électronique pouvant être utilisée sous deux modes différents : chaque individu est l'histoire d'une pièce résumée par un vecteur de dimension 3 contenant « nombre de pannes », « durée de fonctionnement en mode 1 » et « durée de fonctionnement en mode 2 ». Les voici représentés :



Discuter du choix d'un modèle de régression adapté à l'étude de ces données.

SOLUTION. Comme les données sont à valeurs entières, il est logique de considérer un modèle linéaire généralisé poissonnien,

$$Pannes \sim \mathcal{P}(r(\beta_0 + \beta_1 m_1 + \beta_2 m_2)).$$

Au vu de la linéarité de dépendance que semble montrer le graphique, il faudrait prendre un lien identité, $r(x) = x$, plutôt que de laisser le lien logarithme par défaut.

Exercice 4 Dans cet exercice, on pourra préférer donner les modèles dans une paramétrisation où la constante n'apparaît pas, mais ce n'est pas nécessaire.

On fait des campagnes publicitaires pour une certaine lessive et l'on s'intéresse à savoir si elles ont un effet significatif ou non (les campagnes).

Pour voir leur effet, on mesure les ventes en période de campagne et en dehors de ces périodes. Ces ventes sont mesurées dans 10 régions différentes. Pour chaque région on a fait plusieurs mesures en période de campagne et en dehors de telles périodes. On a donc des mesures y_{crk}, $c \in \{0, 1\}$, $1 \leq r \leq 10$.

1. On suppose que les campagnes ont toujours le même effet indépendamment du passé, mais qui peut varier d'une région à l'autre. Proposer pour ces données un modèle linéaire gaussien (donner l'équation).

   SOLUTION. $y_{crk} = \mu_{cr} + u_{crk}, \quad u_{rck} \sim N(0, \sigma^2)$.

2. On ne s'intéresse pas spécifiquement à l'effet de la région.

   (a) Proposer un modèle mixte qui prenne quand même ce facteur en compte (on donnera l'équation du modèle).

   SOLUTION. $y_{crk} = \mu_c + \alpha_{cr} + u_{crk}, \quad \alpha_{rc} \sim N(0, \sigma_1^2), \quad u_{rck} \sim N(0, \sigma^2)$.

   D'autres variantes sont possibles.

   (b) Comparer au modèle précédent en terme de nombre de paramètres.

   SOLUTION. Le premier modèle a 20 paramètres de moyenne et 1 paramètre de variance. Le second a deux paramètres de moyenne et 2 de variance, ce qui fait beaucoup moins.

   (c) Améliorer le modèle mixte de sorte qu'on puisse tester s'il y a plus de disparité entre régions en dehors des campagnes que pendant. Expliciter le test (c.-à-d. H_0).

   SOLUTION. $y_{crk} = \mu_c + \alpha_{cr} + u_{crk}, \quad \alpha_{rc} \sim N(0, \sigma_c^2), \quad u_{rck} \sim N(0, \sigma^2)$.
   où σ_c prend deux valeurs σ_0 ou σ_1 selon qu'il y a campagne ou non.
   Tester H_0 :«$\sigma_0 = \sigma_1$» contre son contraire.

Exercice 5 Soit le modèle habituel $y \sim \mathcal{N}(X\beta^*, \sigma_*^2 I)$. Soit un nouvel échantillon $y' \sim \mathcal{N}(x'\beta^*, \sigma_*^2)$. On suppose que x' a été tiré uniformément sur $\{x_1, ...x_n\}$.

Quelle est la variance de $x'\beta^* - x'\hat{\beta}$? Quelle est la variance de $y' - x'\hat{\beta}$? (Tenir compte de la loi de x' dans le calcul des variances).

Solution. Ces variables sont centrées, il s'agit donc de calculer l'espérance du carré. On peut faire le calcul de plusieurs façons différentes, la clef étant l'indépendance du tirage de x' et des bruits de régression ; on le fait ici avec l'astuce utilisée dans le cours

$$\begin{aligned}
E[(x'(\beta^* - \hat{\beta}))^2] =& E[x'(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T x'^T] \\
=& E[Tr(x'^T x'(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T)] \\
=& TrE[x'^T x'(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T] \\
=& TrE[x'^T x']E[(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T] \\
=& Tr[(n^{-1}\sum x_i^T x_i)(X^T X)^{-1}\sigma_*^2] \\
=& \frac{p}{n}\sigma_*^2.
\end{aligned}$$

Puis $Var(y' - x'\hat{\beta}) = Var(u' + x'\beta^* - x'\hat{\beta}) = Var(u') + Var(x'\beta^* - x'\hat{\beta}) = (1 + p/n)\sigma_*^2$.

Exercice 6 Soit le modèle :

$$y_{ik} = \mu_i + u_{ik}, \quad u_{ik} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, 2, \quad k = 1, ...n_i.$$

1. Donner l'expression de $\hat{\mu}_1$ et $\hat{\mu}_2$.
   Solution. $\hat{\mu}_i = n_i^{-1}\sum_k y_{ik}$.
2. On suppose $\mu_1 = \mu_2$. Quelle est la loi de $\hat{\mu}_1 - \hat{\mu}_2$? (faire le calcul)
   Solution. $\hat{\mu}_1 - \hat{\mu}_2 = \frac{1}{n_1}\sum u_{1i} - \frac{1}{n_2}\sum u_{2i} \sim \mathcal{N}(0, \sigma^2(n_1^{-1} + n_2^{-1}))$
3. (μ_1, μ_2) est à nouveau quelconque. On rappelle que sous des hypothèses adéquates $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$. Une autre estimation du modèle est faite sur des données indépendantes, de même loi que les précédentes, mais en nombre différent $(n'_i \neq n_i)$. On obtient en particulier une estimée $\hat{\sigma}'$ de σ.
   (a) Préciser la loi de $S = (n-2)\hat{\sigma}^2 + (n'-2)(\hat{\sigma}')^2$.
   Solution. On sait que $\sigma^{-2}S$ suit un $\chi^2_{n+n'-4}$ (en raison de l'indépendance des deux estimateurs et du fait qu'il suivent chacun un χ_2).
   (b) Soit la statistique
   $$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{n_1^{-1} + n_2^{-1}}\sqrt{S/(n + n' - 4)}}.$$
   Que permet-elle de tester ? Quelle est sa loi sous H_0 ?
   Solution. Le test permet des décider si les moyennes sont significativement différentes. H_0 correspond à l'identité des moyennes. Sous H_0, le numérateur $\hat{\mu}_1 - \hat{\mu}_2$, est indépendant de $\hat{\sigma}$ et $\hat{\sigma}'$, et donc de S. Par conséquent, en raison des normalisations, sous H_0 la statistique T suit une loi de Student de paramètre $n + n' - 4$.