

- 3.1 On considère trois modèles de régression emboîtés afin de modéliser le nombre d'accidents de voiture selon la région (region). La variable catégorielle niveaux de risque (risque) a 3 niveaux et le nombre d'années d'expérience (expe) au volant est un facteur à 4 niveaux.

Modèle	variables	$p+1$	$\ell(\hat{\beta})$	AIC	BIC
M <sub>1</sub>	risque	3	-244.566	495.132	510.362
M <sub>2</sub>	risque + region	★	-151.620	★	★
M <sub>3</sub>	risque + region + expe	10	-139.734	299.468	350.235

Table 1: Mesures d'adéquation de trois modèles emboîtés avec les covariables, le nombre de coefficients du modèle ( $p+1$ ), la valeur de la log-vraisemblance évaluée au maximum de vraisemblance ( $\ell(\hat{\beta})$ ) et les critères d'information.

Quelle est la différence entre le AIC et le BIC du modèle M<sub>2</sub> (en valeur absolue)?

**Solution**

La différence est approximativement de 35.54. Les seules informations requises ici sont le nombre de paramètres du modèle M<sub>2</sub> et la taille de l'échantillon. On peut utiliser la première ligne pour calculer cette dernière,

$$\text{BIC} = -2\ell(\text{M}_1) + 3\ln(n)$$

et on trouve en arrondissant  $n = 1184$ . Le nombre de paramètres dépend du nombre de niveaux de la variable région. On trouve  $\text{DF}(\text{M}_3) - \text{DF}(\text{M}_1) - 3 = 7$ ; il y a  $K = 4$  catégories pour les années d'expérience, mais seulement trois paramètres additionnels dans le modèle. La différence  $\text{BIC} - \text{AIC} = 7 \cdot \{\ln(1184) - 2\}$ .

- 3.2 Une variable aléatoire  $X$  suit une loi géométrique de paramètre  $p$  si sa fonction de masse est

$$P(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, \dots$$

- Écrivez la vraisemblance et la log-vraisemblance d'un échantillon aléatoire de taille  $n$  si les observations sont indépendantes.
- Dérivez l'estimateur du maximum de vraisemblance pour le paramètre  $p$ .
- Calculez l'information observée.
- Supposons qu'on a un échantillon de 15 observations,  $\{5, 6, 3, 7, 1, 2, 11, 8, 7, 34, 1, 7, 10, 1, 0\}$ , dont la somme est 103. Calculez l'estimé du maximum de vraisemblance et son erreur-type approximative.
- Calculez la statistique du rapport de vraisemblance et la statistique de Wald pour un test à niveau 5% de l'hypothèse  $\mathcal{H}_0 : p_0 = 0.1$  contre l'alternative bilatérale  $\mathcal{H}_a : p_0 \neq 0.1$ .

**Solution**

(a)

$$L(p; \mathbf{x}) = \prod_{i=1}^n (1 - p)^{x_i - 1} p = (1 - p)^{\sum_{i=1}^n (x_i - 1)} p^n$$

$$\ell(p; \mathbf{x}) = \ln(1 - p) \sum_{i=1}^n (x_i - 1) + n \ln(p)$$

(b)

$$\frac{d}{dp} \ell(p; \mathbf{x}) = -\frac{1}{(1 - p)} \sum_{i=1}^n (x_i - 1) + \frac{n}{p}$$

Si on fixe la fonction de score à zéro et qu'on réarrange l'expression, on obtient

$$\frac{1}{p} = \frac{1}{n} \sum_{i=1}^n x_i$$

La dérivée deuxième est négative, donc l'estimateur du maximum de vraisemblance est la réciproque de la moyenne,  $\hat{p} = \bar{X}^{-1}$ .

- (c) La fonction d'information observée vaut  $-1$  fois la hessienne de la log-vraisemblance,

$$-\frac{d^2}{dp^2} \ell(p; \mathbf{x}) = \frac{n(\bar{x} - 1)}{(1 - p)^2} + \frac{n}{p^2}.$$

Si on évalue à l'estimé du maximum de vraisemblance, on obtient  $j(\hat{p}) = n\bar{x}^3/(\bar{x} - 1)$ .

- (d) L'estimé du maximum de vraisemblance est 0.1456 et son erreur-type est 0.0347.  
 (e) La statistique de Wald et la statistique du rapport de vraisemblance valent

$$W = (\hat{p} - p_0)^2 / \text{Va}(\hat{p}) = (0.1456 - 0.1)^2 / 0.0347 = 1.72$$

$$R = 2\{\ell(\hat{p}) - \ell(p_0)\} = 2\{-45.11 + 45.39\} = 0.558.$$

Les deux peuvent être comparées à une loi  $\chi_1^2$ : les valeurs- $p$  sont 0.19 et 0.45 (obtenues à l'aide d'un logiciel), donc on ne rejette pas  $\mathcal{H}_0 : p = 0.1$ .

- 3.3 On considère le temps avant défaillance de machines sur la base de leur niveau de corrosion  $w$ . Spécifiquement, le temps avant défaillance,  $T$ , est modélisé à l'aide d'une loi exponentielle de densité  $f(t) = \lambda \exp(-\lambda t)$ , mais d'intensité  $\lambda = aw^b$ ; si  $b = 0$ , le temps de défaillance moyen est constant et vaut  $E(T_i) = a^{-1}$ . On suppose que les  $n$  observations sont indépendantes et que les niveaux de corrosion  $w_i$  sont supposés connus (donc fixes). [Coles (2001)]

- (a) Écrivez la log-vraisemblance du modèle  
 (b) Dérivez les matrices d'informations observées et de Fisher.  
 (c) Montrez que la log-vraisemblance profilée pour  $b$  est

$$\ell_p(b) = n \ln(\hat{a}_b) + b \sum_{i=1}^n \ln(w_i) - \hat{a}_b \sum_{i=1}^n w_i^b t_i,$$

et dérivez une formule explicite pour l'estimateur du maximum de vraisemblance partiel  $\hat{a}_b$ .

### Solution

- (a) La log-vraisemblance est

$$\ell(a, b; \mathbf{w}, \mathbf{t}) = n \ln(a) + b \sum_{i=1}^n \ln(w_i) - a \sum_{i=1}^n w_i^b t_i$$

- (b) L'information de Fisher est l'espérance du négatif de la hessienne. L'information observée est

$$j(a, b) = - \begin{pmatrix} \partial^2 \ell / \partial a^2 & \partial^2 \ell / \partial a \partial b \\ \partial^2 \ell / \partial b \partial a & \partial^2 \ell / \partial b^2 \end{pmatrix} = \begin{pmatrix} na^{-2} & \sum_{i=1}^n w_i^b t_i \ln(w_i) \\ \sum_{i=1}^n w_i^b t_i \ln(w_i) & a \sum_{i=1}^n w_i^b t_i \ln^2(w_i) \end{pmatrix}.$$

Pour obtenir l'information de Fisher, on calcule l'espérance de chaque entrée de la matrice. Seuls les fonctions des  $T_i$  sont aléatoires et chaque entrée n'a que des facteurs linéaires ( $T_i$  fois une constante). Puisque  $E(T_i) =$

$a^{-1} w_i^{-b}$ , l'information de Fisher est

$$I(a, b) = \begin{pmatrix} na^{-2} & a^{-1} \sum_{i=1}^n \ln(w_i) \\ a^{-1} \sum_{i=1}^n \ln(w_i) & \sum_{i=1}^n \ln^2(w_i) \end{pmatrix}$$

- (c) On obtient l'estimateur du maximum de vraisemblance partiel  $\hat{a}_b$  en dérivant la log-vraisemblance par rapport à  $a$  tout en traitant  $b$  comme fixe,

$$\frac{\partial \ell}{\partial a} = \frac{n}{a} - \sum_{i=1}^n w_i^b t_i = 0$$

et on voit que la valeur de  $a$  qui maximise cette expression est  $\hat{a}_b = n / \sum_{i=1}^n w_i^b t_i$ , puisque la dérivée deuxième calculée précédemment est négative.

On considère un modèle pour le nombre de ventes quotidiennes dans un magasin, supposées indépendantes. Votre gestionnaire vous indique que ce dernier fluctue selon qu'il y ait des soldes ou pas. Vous spécifiez un modèle de Poisson avec fonction de masse

$$P(Y_i = y_i | \text{soldes}_i) = \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, \dots$$

et on modélise  $\lambda_i = \exp(\beta_0 + \beta_1 \text{soldes}_i)$ , où  $\text{soldes}_i$  est un indicateur binaire qui vaut un lors des soldes et zéro autrement.

- (a) Dérivez l'estimateur du maximum de vraisemblance pour  $(\beta_0, \beta_1)$ . *Indice: les estimateurs du maximum de vraisemblance sont invariants aux reparamétrisations. Considérez les deux échantillons solde/hors solde.*

### Solution

On utilise l'indice  $s$  pour dénoter les événements correspondants à  $\text{soldes} = 1$  et  $r$  si  $\text{soldes} = 0$ . Le nombre d'observations totales, hors-soldes et pendant les soldes sont respectivement dénotées  $n = 12$ ,  $n_r = 7$ ,  $n_s = 5$ . Soit  $\lambda_s$  ( $\lambda_r$ ) l'espérance du nombre de ventes quotidiennes pendant (et hors) soldes et  $\bar{y}_s$ ,  $\bar{y}_r$  et  $\bar{y}$  la moyenne du nombre de ventes quotidiennes quand  $\text{soldes} = 0$ ,  $\text{soldes} = 1$  et la moyenne empirique de l'échantillon complet. La façon la plus simple de dériver l'estimateur du maximum de vraisemblance (EMV) est d'utiliser l'invariance aux transformations monotones. On sait que  $\hat{\lambda}_r$  et  $\hat{\lambda}_s$  correspondent aux moyennes empiriques des ventes pour les différentes périodes, d'où

$$\hat{\lambda}_r = \exp(\hat{\beta}_0), \quad \hat{\lambda}_s = \exp(\hat{\beta}_0 + \hat{\beta}_1).$$

On obtient donc  $\hat{\beta}_0 = \ln(\bar{y}_r)$  et  $\hat{\beta}_1 = \ln(\bar{y}_s / \bar{y}_r)$ .

Une façon plus directe (et pénible) d'obtenir  $\hat{\beta}$  est de différencier la log vraisemblance, égaliser le gradient à

zéro et résoudre simultanément pour  $\hat{\beta}_0$  et  $\hat{\beta}_1$ .

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^n y_i(\beta_0 + \beta_1 \text{solde}_i) - \exp(\beta_0) \sum_{i=1}^n \exp(\beta_1 \text{solde}_i) - \sum_{i=1}^n \ln(y_i!) \\ \frac{\partial \ell}{\partial \beta_0} &= \sum_{i=1}^n y_i - \exp(\beta_0) \sum_{i=1}^n \exp(\beta_1 \text{solde}_i) = n\bar{y} - \exp(\beta_0)\{n_r + n_s \exp(\beta_1)\} \\ \frac{\partial \ell}{\partial \beta_1} &= \sum_{i=1}^n y_i \text{solde}_i - \sum_{i=1}^n \text{solde}_i \exp(\beta_0 + \beta_1 \text{solde}_i) = n_s \bar{y}_s - \exp(\beta_0) n_s \exp(\beta_1) \\ \frac{\partial^2 \ell}{\partial \beta_0^2} &= - \sum_{i=1}^n \exp(\beta_0 + \beta_1 \text{solde}_i) \\ \frac{\partial^2 \ell}{\partial \beta_1^2} &= - \sum_{i=1}^n \text{solde}_i \exp(\beta_0 + \beta_1 \text{solde}_i) = -n_s \exp(\beta_0 + \beta_1) \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} &= -n_s \exp(\beta_0 + \beta_1)\end{aligned}$$

Les EMVS satisfont  $\exp(\beta_0) = \bar{y}_s / \exp(\beta_1)$ ; en substituant ce terme dans l'expression pour  $\partial \ell / \partial \beta_0 = 0$  avec  $n\bar{y} = n_r \bar{y}_r + n_s \bar{y}_s$ , on obtient

$$n\bar{y} \exp(\beta_1) = \bar{y}_s \{n_r + n_s \exp(\beta_1)\}$$

et un réarrangement des différents termes donne

$$\frac{n_r \bar{y}_s}{n\bar{y} - n_s \bar{y}_s} = \frac{\bar{y}_s}{\bar{y}_r} = \exp(\beta_1).$$

- (b) Calculez les estimés si on a un échantillon du nombre de ventes pour 12 jours indépendants, soit {2; 5; 9; 3; 6; 7; 11} hors soldes, et {12; 9; 10; 9; 7} pendant les soldes.

**Solution**

À l'aide des formules ou d'un logiciel, on obtient  $\hat{\beta}_0 = 1.8153$  et  $\hat{\beta}_1 = 0.4254$ .

- (c) Calculez la matrice d'information observée et utilisez-la pour dériver les erreurs-type de  $\hat{\beta}_0$  et de  $\hat{\beta}_1$  et un intervalle de confiance à niveau 95%.

**Solution**

L'expression pour la hessienne se simplifie si on substitue  $\boldsymbol{\beta}$  par  $\hat{\boldsymbol{\beta}}$ : on obtient

$$j(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} n\bar{y} & n_s \bar{y}_s \\ n_s \bar{y}_s & n_s \bar{y}_s \end{pmatrix} = \begin{pmatrix} 90 & 47 \\ 47 & 47 \end{pmatrix}$$

En inversant la matrice  $2 \times 2$  et en prenant la racine carré des termes sur la diagonale de  $j^{-1}(\hat{\boldsymbol{\beta}})$ , on obtient les erreurs-type de  $\hat{\boldsymbol{\beta}}$ , à savoir  $\text{se}(\hat{\beta}_0) = 43^{-1/2} = 0.1525$  et  $\text{se}(\hat{\beta}_1) = (90/43 \cdot 47)^{1/2} = 0.2110$ . L'intervalle de confiance à 95% pour  $\beta_0$  et  $\beta_1$  sont respectivement [1.5164; 2.1142] et [0.0118; 0.839].

- (d) Votre gérant veut savoir si les profits quotidiens moyens pendant les soldes sont différents des jours ordinaires, sachant que le profit moyen pour les jours hors solde est de 25\$ par transaction, mais seulement 20\$ pendant les soldes. Faites un test de rapport de vraisemblance (profilée) pour estimer cette hypothèse. *Indice: écrivez l'hypothèse nulle en fonction des paramètres du modèles  $\beta_0$  et  $\beta_1$  (difficile).*

**Solution**

On veut tester si les profits quotidiens sont les même pendant les soldes que en période régulière, ce qui

revient à  $\mathcal{H}_0 : 20 \exp(\beta_0 + \beta_1) = 25 \exp(\beta_0)$ , ou  $\beta_1 = \log(5/4)$ . On peut donc calculer le maximum de vraisemblance contraint,

$$\ell_p(\beta_0) = \sum_{i=1}^n y_i \{\beta_0 + \log(5/4) \text{solde}_i\} - \exp(\beta_0) \sum_{i=1}^n (5/4)^{\text{sale}_i}$$

En calculant la dérivée première et en fixant cette dérivée à zéro, on obtient l'égalité  $n\bar{y} - \exp(\beta_0)(n_r + 5n_s/4) = 0$  et donc  $\hat{\beta}_0 = \ln(n\bar{y}) - \ln(n_r + 5n_s/4) = 1.9158$ . On peut calculer le maximum de log vraisemblance sous l'hypothèse nulle et alternative,  $\ell(\hat{\beta}) = 93.37082$  et  $\ell_p(\hat{\beta}_0) = 92.91$ . La statistique du test de rapport de vraisemblance vaut  $R = 0.92$  et on compare cette valeur à une loi  $\chi_1^2$  dont le 0.95 quantile est 3.84; on ne rejette pas l'hypothèse bilatérale que les profits sont les mêmes peu importe s'il y a des soldes ou pas; la valeur- $p$  est 0.337.