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	$p+1$	$\ell(\widehat{\boldsymbol{\beta}})$	AIC	BIC
M_1	`risque`	3	−244.566	495.132	510.362
M_2	`risque` + `region`	⋆	−151.620	⋆	⋆
M_3	`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 ($\ell(\widehat{\boldsymbol{\beta}})$) et les critères d'information.

Quelle est la différence entre le AIC et le BIC du modèle M_2 (en valeur absolue)?

Solution

La différence est approximativement de 35.54. Les seules informations requises ici sont le nombre de paramètres du modèle M_2 et la taille de l'échantillon. On peut utiliser la première ligne pour calculer cette dernière,

$$\mathsf{BIC} = -2\ell(M_1) + 3\ln(n)$$

et on trouve en arrondissant $n = 1184$. Le nombre de paramètres dépend du nombre de niveaux de la variable région. On trouve $\mathrm{DF}(M_3) - \mathrm{DF}(M_1) - 3 = 7$; il y $K = 4$ catégories pour les années d'expérience, mais seulement trois paramètres additionnels dans le modèle. La différence $\mathsf{BIC} - \mathsf{AIC} = 7 \cdot \{\ln(1184) - 2\}$.

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

$$\mathsf{P}(X = x) = (1-p)^{x-1}p, \qquad x = 1, 2, \ldots$$

(a) Écrivez la vraisemblance et la log-vraisemblance d'un échantillon aléatoire de taille n si les observations sont indépendantes.
(b) Dérivez l'estimateur du maximum de vraisemblance pour le paramètre p.
(c) Calculez l'information observée.
(d) 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.
(e) Calculez la statistique du rapport de vraisemblance et la statistique de Wald pour un test à niveau 5% de l'hypothèse $\mathscr{H}_0 : p_0 = 0.1$ contre l'alternative bilatérale $\mathscr{H}_a : p_0 \neq 0.1$.

Solution

(a)

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

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

(b)

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

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, $\widehat{p} = \overline{X}^{-1}$.

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

$$-\frac{\mathrm{d}^2}{\mathrm{d}p^2}\ell(p;\boldsymbol{x}) = \frac{n(\overline{x}-1)}{(1-p)^2} + \frac{n}{p^2}.$$

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

(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

$$W = (\widehat{p}-p_0)^2/\mathsf{Va}(\widehat{p}) = (0.1456-0.1)^2/0.0347 = 1.72$$
$$R = 2\{\ell(\widehat{p}) - \ell(p_0)\} = 2\{-45.11+45.39\} = 0.558.$$

Les deux peuvent être comparées à une loi χ^2_1: les valeurs-p sont 0.19 et 0.45 (obtenues à l'aide d'un logiciel), donc on ne rejette pas $\mathscr{H}_0 : p = 0.1$.

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

$$\mathsf{P}\left(Y_i = y_i \mid \texttt{soldes}_i\right) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, \ldots$$

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

(a) Dérivez l'estimateur du maximum de vraisemblance pour (β_0, β_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 `solde` = 1 et r si `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 λ_s (λ_r) l'espérance du nombre de ventes quotidiennes pendant (et hors) soldes et $\overline{y}_s$, $\overline{y}_r$ et $\overline{y}$ la moyenne du nombre de ventes quotidiennes quand `solde` = 0, `solde` = 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 $\widehat{\lambda}_r$ et $\widehat{\lambda}_s$ correspondent aux moyennes empiriques des ventes pour les différentes périodes, d'où

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

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

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

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

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

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

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

et un réarrangement des différents termes donne

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

(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 $\widehat{\beta}_0 = 1.8153$ et $\widehat{\beta}_1 = 0.4254$.

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

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

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

En inversant la matrice 2×2 et en prenant la racine carré des termes sur la diagonale de $j^{-1}(\widehat{\boldsymbol{\beta}})$, on obtient les erreurs-type de $\widehat{\boldsymbol{\beta}}$, à savoir $\text{se}(\widehat{\beta}_0) = 43^{-1/2} = 0.1525$ et $\text{se}(\widehat{\beta}_1) = (90/43\cdot 47)^{1/2} = 0.2110$. L'intervalle de confiance à 95% pour β_0 et β_1 sont respectivemement $[1.5164;2.1142]$ 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 β_0 et β_1* (**difficile**).

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

Compilé le 16/12/2020 à 18:57

revient à $\mathscr{H}_0 : 20\exp(\beta_0+\beta_1) = 25\exp(\beta_0)$, ou $\beta_1 = \ln(5/4)$. On peut tester cette hypothèse en regardant si $\ln(5/4)$ fait partie de l'intervalle de confiance pour β_1.
Pour obtenir une valeur-p, il faut procéder autrement. On peut donc calculer le maximum de vraisemblance contraint,

$$\ell_{\mathrm{p}}(\beta_0) = \sum_{i=1}^{n} y_i\{\beta_0 + \ln(5/4)\texttt{solde}_i\} - \exp(\beta_0)\sum_{i=1}^{n}(5/4)^{\texttt{sale}_i}$$

En calculant la dérivée première et en fixant cette dérivée à zéro, on obtient l'égalité $n\overline{y} - \exp(\beta_0)(n_r + 5n_s/4) = 0$ et donc $\widehat{\beta}_0 = \ln(n\overline{y}) - \ln(n_r + 5n_s/4) = 1.9158$. On peut calculer le maximum de log vraisemblance sous l'hypothèse nulle et alternative, $\ell(\widehat{\boldsymbol{\beta}}) = 93.37082$ et $\ell_{\mathrm{p}}(\widehat{\beta}_0) = 92.91$. La statistique du test de rapport de vraisemblance vaut $R = 0.92$ et on compare cette valeur à une loi χ^2_1 dont le 0.95 quantile est 3.84; on ne rejette pas l'hypothèse bilatérale que les profits sont les même peut importe s'il y a des soldes ou pas; la valeur-p est 0.337.

Compilé le 16/12/2020 à 18:57