4.1 **Régression logistique :** on modélise le salaire de professeurs d'une université américaine pour une période de neuf mois. Le jeu de données `salaireprof` contient les variables suivantes :

— `sexe` : sexe, soit homme (0) ou femme (1);
— `echelon` : variable catégorielle, soit adjoint(e) (1), soit aggrégé(e) (2), soit titulaire (3);
— `diplome` : diplôme le plus élevé complété, soit maîtrise (0) ou doctorat (1);
— `anec` : nombre d'années au sein de l'échelon académique;
— `andi` : nombre d'années depuis l'obtention dernier diplôme;
— `salaire` : salaire sur neuf mois (en dollars américains).

(a) Ajustez un modèle logistique pour modéliser la probabilité qu'un professeur ait un salaire supérieur à 105 000 USD en fonction de `diplome`, `sexe`, `anec` and `andi`. Écrivez l'équation du modèle ajusté et interprétez les paramètres du modèle.

Solution

L'équation du modèle ajusté est

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\beta_0 + \beta_1 \texttt{diplome}_i + \beta_2 \texttt{sexe}_i + \beta_3 \texttt{anec}_i + \beta_4 \texttt{andi}_i\right)$$

— $\text{expit}(\beta_0)$ est la probabilité qu'un nouveau professeur adjoint qui termine sa maîtrise gagne plus de 105 000 dollars américains. L'estimation de cette probabilité est de 0,000286.
— $\widehat{\beta}_1 = 18,58$; la cote d'un professeur avec un doctorat est 18,58 fois plus élevée que pour un professeur avec une maîtrise, toute chose étant égale par ailleurs.
— $\widehat{\beta}_2 = 0,30$; la cote des femmes est $0,3$ fois celles des hommes (70% moindre), ceteris paribus.
— Les deux derniers coefficients ne peuvent être interprétés séparéments, à moins que la personne ne change d'échelon académique (auquel cas `anec` passe de x à 1. En général, la cote d'un professeur qui demeure au même rang augmente de $\exp(\widehat{\beta}_3 + \widehat{\beta}_4) = \exp(1,276 + 1,171) = 11,55$ par année.

(b) Ajoutez la covariable `echelon` en plus des autres variables. Arrivez-vous à identifier un problème avec ce nouveau modèle? Si oui, expliquez les résultats.

Solution

L'information fournie par l'`echelon` et le nombre d'années depuis le diplôme et au sein de l'échelon est partiellement redondante et engendre de la collinéarité. L'ordonnée à l'origine estimée est $\widehat{\beta}_0 = -25,3$ dans R /$-11,1$ dans SAS , le coefficient pour les professeurs aggrégés $\widehat{\beta}_{\texttt{aggrege}} = 0,46$ dans R /$-5,72$ dans SAS et celui pour les titulaires de $\widehat{\beta}_{\texttt{titulaire}} = 26,1$ dans R /$11,92$ dans SAS; cela signifie que le modèle prédit que tout le monde qui est dans les deux premiers échelons (adjoints et aggrégés) a (en pratique) une probabilité nulle d'avoir un salaire excédant 105 000 dollars américains. Ces résultats numériques et l'instabilité des estimés sont dus à à la quasi-séparation de variables.

4.2 Un chercheur en pédagogie est intéressé par la relation entre le nombre de prix remportés par des élèves d'une école secondaire et leurs notes en mathématique. On considère également le type de programme que les élèves fréquentent. Les données `prix` contiennent les variables suivantes :

— `nprix` : variable réponse, décompte du nombre total de prix reçus au cours de l'année scolaire.
— `math` : note de l'étudiant à l'examen final de mathématiques
— `prog` : programme d'étude de l'étudiant(e), un choix parmi général (1), programme enrichi (2) ou professionnel (3).

Ajustez une régression de Poisson et une régression binomiale négative en fonction des covariables `math` et `prog`. Comparez les résultats des deux modèles, indiquez si l'un ou l'autre est adéquat.

Solution

Le test du rapport de vraisemblance pour $\mathscr{H}_0 : k = 0$ (paramètre de dispersion du modèle de régression binomiale négative) permet de comparer le modèle de Poisson (sous $\mathscr{H}_0$) au modèle binomiale négative (sous $\mathscr{H}_1$). La valeur-p est $0,096$; on ne rejette pas l'hypothèse nulle ; l'estimé du coefficient de dispersion est $\widehat{k} = 0,1635 = 1/6,114$. Le ratio de la déviance par rapport au degrés de liberté résiduels du modèle de Poisson est $0,97$, ce qui suggère que le modèle est adéquat.

4.3 **Taux** : les données `enfantsfiji` contiennent des information sur le nombre d'enfants nés, tirées de l'Étude de fertilité des Fiji. Les variables suivantes ont été mesurées pour plusieurs groupes de femmes :
- `nfemmes` : nombre de femmes dans le groupe ;
- `nenfants` : variable réponse, nombre d'enfants nés ;
- `dur` : temps (en années) depuis le mariage, un parmi 0–4 (1), 5–9 (2), 10–14 (3), 15–19 (4), 20–24 (5) et plus de 25 ans (6).
- `res` : variable catégorielle pour résidence, une parmi Suva (1), région urbaine (2) ou région rurale (3).
- `educ` : variable ordinale indiquant le niveau d'éducation, un parmi aucun (1), début du primaire (2), fin du primaire (3), secondaire ou plus (4).
- `var` : variance estimée intra-groupe du nombre d'enfants nés

(a) Tracez un graphique du nombre d'enfants nés (`nenfants`) en fonction du nombre de femmes dans le groupe (`nfemmes`) et commentez.

Solution

Il semble y avoir une relation linéaire claire entre les deux variables à l'échelle logarithmique. La relation entre moyenne et variance est nonlinéaire ; la variabilité augmente pour les plus grands dénombrements.

- Devrait-on inclure un terme de décalage ? Justifiez votre réponse.
- Si on inclut pas de décalage, quelle fonction (si aucune) de `nfemmes` devrait être incluse comme prédicteur ?
- Si on inclut un décalage, comment ce modèle se compare-t-il au modèle qui inclut log(`nfemmes`) comme prédicteur ?

Solution

Les dénombrements ne sont pas comparables, donc un décalage est utile d'office. Le terme à inclure dans la moyenne pour obtenir le taux de naissances est log(`nfemmes`). Si on voulait considérer plutôt un terme supplémentaire dans la moyenne, c'est log(`nfemmes`) qui semble le plus adéquat au vu du graphe. On peut vérifier si le décalage est adéquat en testant si le coefficient pourrait être égal à un ; l'intervalle de confiance à 95% basé sur la vraisemblance profilée est $[0,97; 1,06]$, ce qui ne suggère rien contre l'inclusion du nombre de femmes comme terme de décalage.

(b) Ajustez un modèle de régression de Poisson avec un décalage en incluant les trois variables catégorielles `dur`, `res` et `educ` sans interactions. Lequel des trois prédicteurs est le plus significatif ?

Solution

La durée du mariage est la variable la plus significative globalement ; la statistique est plus grande, mais il faut ajuster pour le nombre de paramètres additionnels (cinq) ; la valeur-p est la plus petite.

(c) Interprétez les estimés des coefficients du modèle ajusté.

Solution

L'interprétation change selon la paramétrisation du modèle ; la solution suivante utilise les plus faibles catégories comme niveaux de référence, soit $0 - -4$ ans depuie le marriage, habitantes de l'île de Suva, sans éducation).

- Le nombre moyen d'enfant nés par femmes pour les catégories de référence est $\exp(\widehat{\beta}_0) = 0,89$
- Le nombre d'enfants par femme augmente par un facteur $2,7$ pour 5–9 ans de mariage(respectivement 3,93 pour 10–14 ans ; 5,02 pour 15-19 ans ; 5,96 pour 20-24 ans et finalement 7,2 pour 25 ans et plus) par rapport à 0–4 ans de mariage,, *ceteris paribus*.

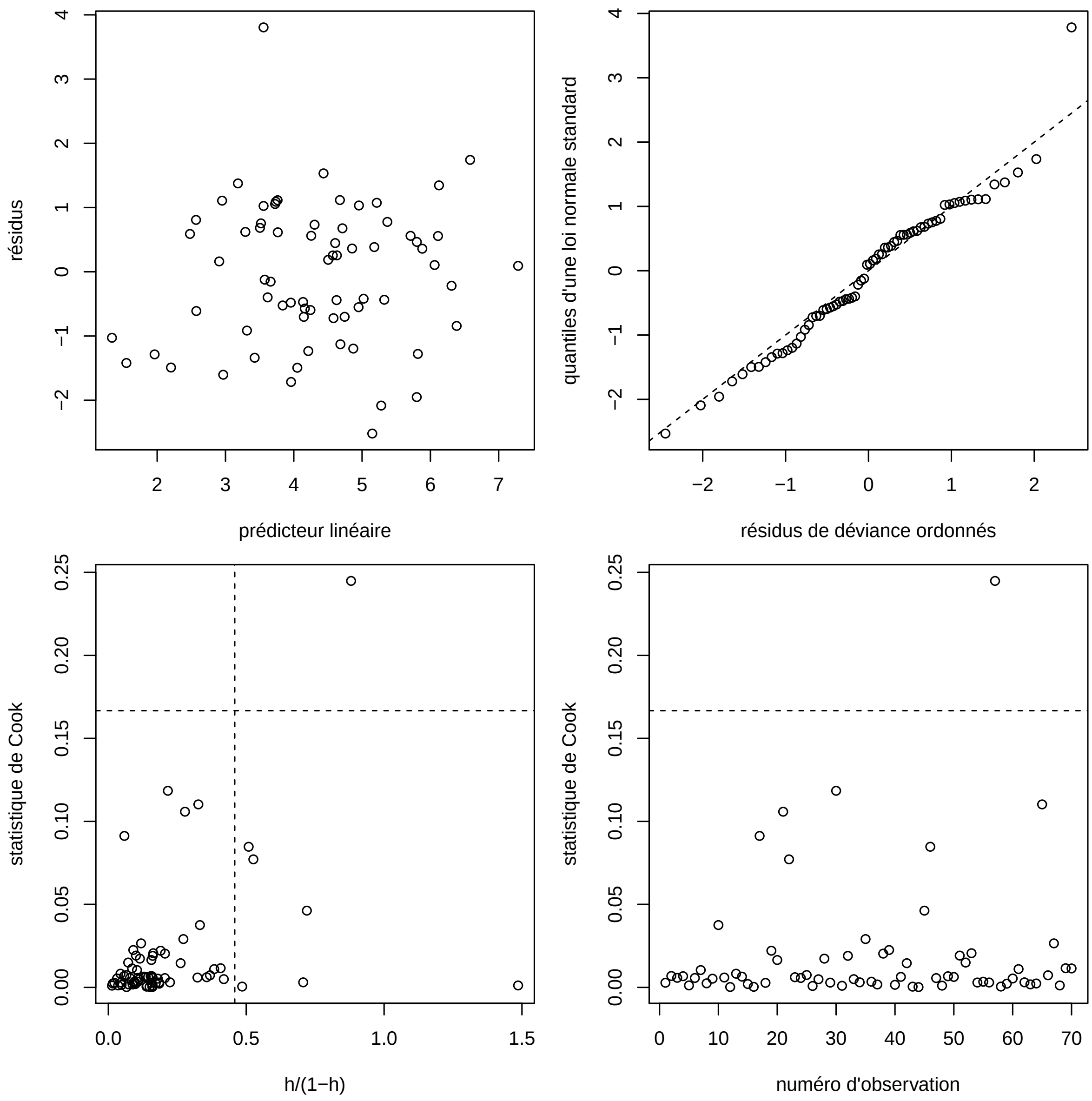


FIGURE 1 – Nombre d'enfants nés en fonction du nombre de femmes par groupe sur l'échelle log-log (gauche) et moyenne versus variance du nombre d'enfants nés par groupe (droite).

— L'augmentation relative du taux de natalité pour les femmes ayant complété au plus début-primaire est de 1,02 (respectivement 0,9 pour éducation fin-primaire et 0,73 pour une éducation secondaire ou supérieure), par rapport à aucune éducation. Règle générale, plus les femmes ont étudié, moins elles ont d'enfants.

— Les personnes des zones urbains aux îles Fiji on 1,12 fois plus d'enfants qu'à Suva, tandis que ceux des zones rurales ont 1,16 fois plus qu'à Suva, pour un même niveau d'éducation et la même tranche d'années depuis le mariage.

(d) Déterminez si une interaction entre `educ` et `duration` est utile en faisant un test du rapport de vraisemblance.

Solution
L'ajout de l'interaction donne 15 paramètres additionnels. On peut regarder le tableau de type 3 ou comparer la déviance du modèle avec et sans l'interaction. La statistique du rapport de vraisemblance est 15,86 ; a valeur-p de l'hypothèse nulle qu'ils sont tous simultanément zéro est 0,3912. On conclut qu'il n'y a pas de preuves que l'ajustement du modèle avec l'interaction est significativement meilleure.

(e) Il est possible de vérifier l'adéquation du modèle à l'aide de diagnostics graphiques, notamment en étudiant les résidus de déviance du modèle de Poisson. [1]Produisez des diagnostics graphiques de (a) prédicteur linéaire versus résidus de déviance (b) diagramme quantile-quantile des résidus de déviance, (c) effet levier et (d)distance Cook en fonction des observations. Commentez sur l'ajustement du modèle de Poisson qui inclut trois variables explicatives catégorielles et le décalage. *Indication : l'interprétation est semblable à celle des modèles linéaires. Avec SAS , utilisez les options*

```
plots=(resdev(xbeta) leverage cooksd)
```

Le diagramme quantile-quantile peut être produit avec la procédure `univariate`.
Dans R , la fonction `boot::glm.diag.plots` *permet de produire les diagnostics graphiques.*

Solution
Les diagnostics de la Figure 2 semblent bons ; un seul résidu a un effet de levier, mais l'adéquation du modèle est excellente règle générale.

4.4 **Données de location** **`bixiuni`** : BIXI est une entreprise de location de vélos basée à Montréal. Nous avons extrait les données de location de BIXI pour la période 2017-2019 à la station Édouard-Montpetit en face de HEC. Notre intérêt est d'expliquer la variabilité observée dans le nombre de locations quotidiennes de vélos (mesuré par le nombre d'utilisateurs quotidiens) à cette station en fonction du jour de la semaine et d'indicateurs météorologiques. Les variables contenues dans la base de données sont les suivantes :

— `nutilisateurs` : nombre d'utilisateurs quotidiens à la station Édouard-Montpetit.
— `temp` : température (en degrés Celsius)
— `humid` : pourcentage d'humidité relative, prenant des valeurs comprises entre 0 et 100.
— `jour` : variable catégorielle indiquant le jour de la semaine et prenant des valeurs entre dimanche (`1`) et samedi (`7`).
— `fds` : variable binaire valant zéro si la location est effectuée pendant une fin de semaine (samedi ou dimanche) et un sinon.

On considère les quatre modèles suivants pour expliquer `nutilisateurs` :

— modèle 8.4.1 : modèle de régression de Poisson avec la covariable `fds`.
— modèle 8.4.2 : modèle de régression de Poisson, en incluant les covariables `fds`, `humid` et `temp`.
— modèle 8.4.3 : modèle de régression binomiale négative incluant les covariables `fds`, `humid` et `temp`.

1. Règle général, ces diagnostics sont plus difficiles à interpréter que ceux des modèles de régression linéaire parce que les observations sont discrètes tandis que les moyennes ajustées sont continues.

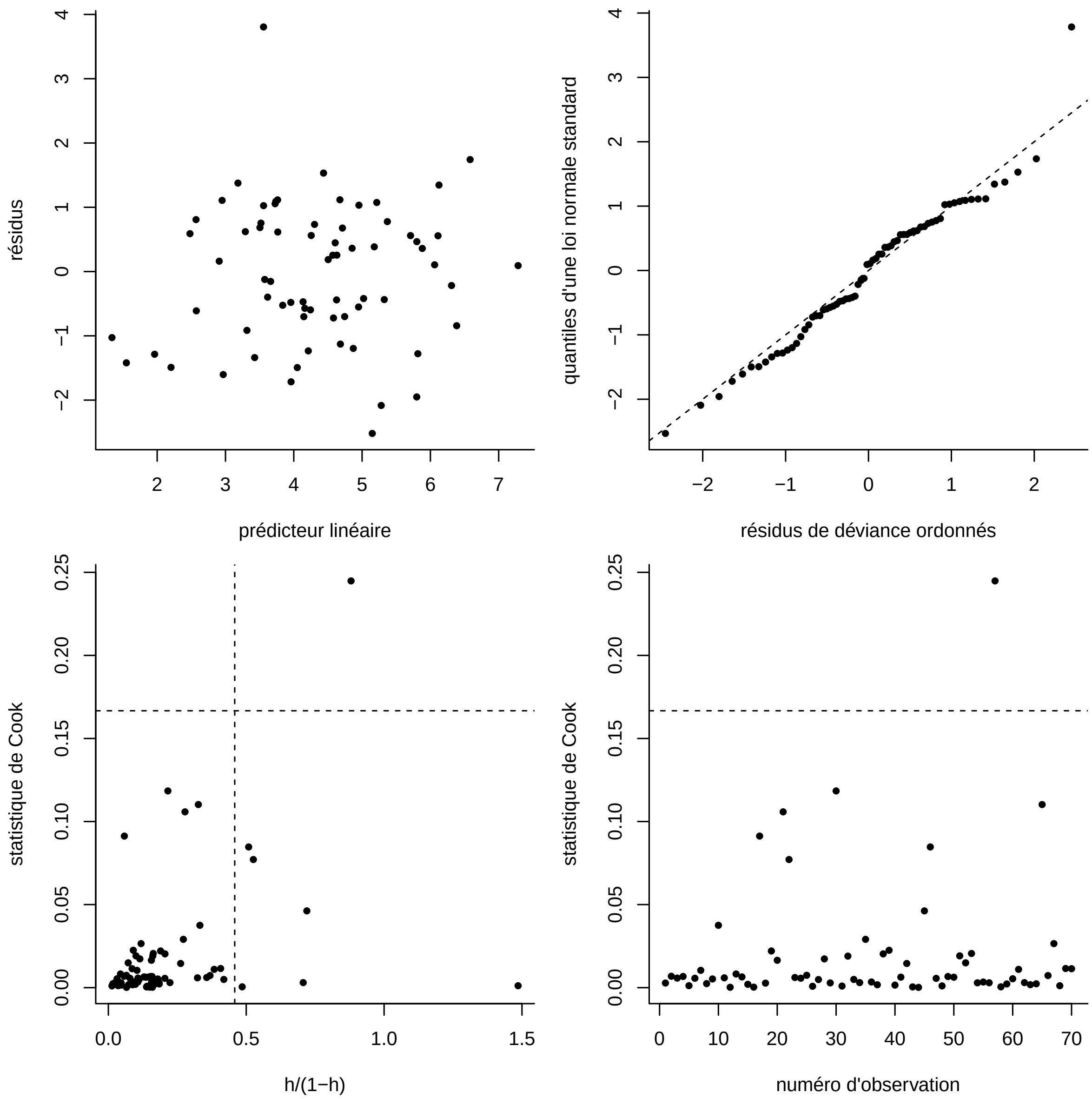


FIGURE 2 – Diagnostics graphiques pour les données `enfantsfiji` : des valeurs ajustés (prédicteur linéaire) contre les résidus de déviance (coin supérieur gauche), diagramme quantile-quantile des résidus de déviance (coin supérieur droit), distance de Cook versus effet de levier (coin inférieur gauche) et distance de Cook en fonction des indices des observations (coin inférieur droit)

— modèle 8.4.4 : modèle de régression binomiale négative incluant les variables explicatives `jour` (catégorielle), `humid` et `temp`.

(a) Est-ce que le modèle 8.4.1 représente une simplification adéquate du modèle 8.4.2 ? Justifiez votre réponse en faisant un test d'hypothèse adéquat.

Solution

Les modèles sont emboîtés, donc on peut utiliser un test du rapport de vraisemblance pour les comparer. L'hypothèse nulle est $\mathscr{H}_0 : \beta_{\texttt{temp}} = \beta_{\texttt{humid}} = 0$. La statistique du test du rapport de vraisemblance est $2 \times (2577{,}3604 - 2190{,}8777) = 772{,}97$, à comparer à une loi χ^2_2. On conclut qu'au moins un des effets linéaires des variables explicatives `humid` et `temp` est statistiquement significatif.

(b) Interprétez le coefficient estimé de l'ordonnée à l'origine et celui de la variable `humid` dans le modèle 8.4.2.

Solution

— Quand l'humidité relative est nulle et que la température est de $0°C$, le nombre moyen d'usagers quotidiens la semaine est $\exp(\widehat{\beta}_0) = 13,07$.

— Pour chaque augmentation de l'humidité relative de 1%, *ceteris paribus*, le nombre moyen d'usagers est multiplié par un facteur $\exp(\widehat{\beta}_{\texttt{humid}}) = \exp(-0,0066) = 0,9934217$, soit une diminution de $0,657\%$.

(c) Quel modèle choisiriez-vous parmi les deux modèles 8.4.2 et 8.4.3 ? Justifiez adéquatement votre réponse en utilisant tous les critères suivants : (a) la déviance (b) le test du rapport de vraisemblance et (c) les critères d'information.

Solution

— statistique de déviance : le rapport de la déviance sur les degrés de liberté est $1954/496 = 3,94$ pour la régression de Poisson tandis que ce même rapport pour le modèle de régression binomiale négative est $522,3013/496 = 1,0530$. Les deux modèles sont comparés aux modèles saturés à l'aide de rapports de vraisemblance ; le modèle binomiale négative est adéquat (le rapport devrait être approximativement un).

— Le test du rapport de vraisemblance (problème irrégulier) entre le modèle 8.4.2 et le modèle 8.4.3 n'est pas donné dans la sortie. La log-vraisemblance complète est $-1808,0756$ pour la régression binomiale négative et $-2190,8777$ pour la régression de Poisson. La statistique du rapport de vraisemblance est égale à deux fois la différence de ces deux quantités, soit, $765,6042$. Si on compare la statistique à $\frac{1}{2}\chi^2_1$, il apparaît que le modèle de Poisson n'est pas une simplification adéquate à cause de la surdispersion.

— Critères d'information : la valeur du AIC (respectivement du BIC) pour la régression de Poisson est $4389,75$ ($4406,6137$), contre $3626,27$ (3647.22) pour la régression binomiale négative ; cela suggère que ce dernier modèle est préférable.

(d) On suppose qu'on aimerait utiliser la variable `jour` à la place de la variable `fds`, comme variable explicative dans le modèle 8.4.8.4. Quelle serait la différence principale si on traitait la variable explicative `jour` comme entière plutôt que catégorielle ? Indiquez laquelle de ces deux possibilités est plus logique dans ce contexte.

Solution

Seule la variable catégorielle est logique dans le contexte. Si on laisse les valeurs entières, cela implique un effet linéaire mais le fait qu'on assigne 1 à dimanche est arbitraire et les jours change de façon cyclique, ce qui ne peut être capturé avec une traîne linéaire.

(e) L'espérance du nombre d'utilisateur selon le modèle 8.4.4 est

$$\begin{aligned}\mathsf{E}\,(\texttt{nutilisateurs}) = \exp\big(&\beta_0 + \beta_1\texttt{temp} + \beta_2\texttt{humid} + \beta_3\mathbf{1}_{\texttt{jour=2}} + \beta_4\mathbf{1}_{\texttt{jour=3}}\\ &+ \beta_5\mathbf{1}_{\texttt{jour=4}} + \beta_6\mathbf{1}_{\texttt{jour=5}} + \beta_7\mathbf{1}_{\texttt{jour=6}} + \beta_8\mathbf{1}_{\texttt{jour=7}}\big).\end{aligned}$$

Écrivez l'hypothèse nulle du test comparant le modèle 8.4.3 au modèle 8.4.4 en fonction des paramètres $\boldsymbol{\beta}$ du modèle, ce faisant démontrant que les deux modèles sont emboîtés. Est-ce que le nombre d'utilisateurs change selon le jour de la semaine ou de fin de semaine ?

Solution

Un énième test de rapport de vraisemblance pour comparer le modèle 8.4.3 et le modèle 8.4.4. L'hypothèse nulle (en fonction des paramètres du modèle) est $\mathscr{H}_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7, \beta_8 = 0$, et donc les modèles sont bel et bien emboîtés. La statistique vaut $2 \times (1808,0756 - 1801,7758) = 12,6$, à comparer à une variable de loi χ^2_5. La valeur-p est $0,027$, on rejette donc $\mathscr{H}_0$ à niveau 5% pour conclure que le nombre moyen (sachant les facteurs météorologiques) varient selon les jours de semaine et de fin de semaine.

4.5 **Tableau de contingence à deux facteurs :** les données de dénombrement sont souvent fournies sous forme de tableaux de contingence ; on considère un tableau bidimensionnel avec J et K niveaux. Le même format peut être utilisé pour stocker le nombre de réussites et d'échecs dans chaque cellule. La moyenne du modèle **saturé** pour la cellule j,k (modèle avec deux effets principaux et une interaction) est

$$\text{logit}(p_{jk}) = \alpha + \beta_j + \gamma_k + \nu_{jk}, \qquad j = 1, \ldots, J-1; k = 1, \ldots, K-1. \tag{M_{s}}$$

et a $JK = 1 + (J-1) + (K-1) + (J-1)(K-1)$ paramètres. On peut considérer des modèles plus simples :

— M_0 : le modèle nul $\text{logit}(p_{jk}) = \alpha$ a un paramètre

— M_1 : le modèle avec uniquement première variable catégorielle, $\text{logit}(p_{jk}) = \alpha + \beta_j (j = 1, \ldots, J-1)$, a J paramètres

— M_2 : le modèle avec uniquement deuxième variable catégorielle, $\text{logit}(p_{jk}) = \alpha + \gamma_k (k = 1, \ldots, K-1)$, a K paramètres

— M_3 : le modèle additif avec les deux effets principaux, $\text{logit}(p_{jk}) = \alpha + \beta_j + \gamma_k (j = 1, \ldots, J-1; k = 1, \ldots, K-1)$, a $J+K-1$ paramètres.

La déviance mesure la **différence d'ajustement** entre le modèle saturé et un modèle emboîté plus simple. Sous des conditions de régularité et en assumant que le nombre d'observations dans chacune des JK cellules tend vers ∞,

$$D(\widehat{\boldsymbol{\beta}}_{M_i}) = 2\{\ell(\widehat{\boldsymbol{\beta}}_{M_{\text{s}}}) - \ell(\widehat{\boldsymbol{\beta}}_{M_i})\} \dot{\sim} \chi^2_{JK-p_i}$$

sous l'hypothèse nulle que le modèle M_i avec p_i paramètres est une simplification adéquate du modèle saturé. Comme pour l'ANOVA, on procède par élimination en partant du modèle le plus compliqué et on compare la différence de défiance entre modèles emboîtés $M_i \subset M_j$; cette différence $D(\widehat{\boldsymbol{\beta}}_{M_i}) - D(\widehat{\boldsymbol{\beta}}_{M_j})$ suit approximativement une loi $\chi^2_{p_j - p_i}$ dans de grands échantillons si M_i est une simplification adéquate de M_j (la comparaison de déviance revient à calculer la statistique du rapport de vraisemblance).

Une fois qu'on a terminé la sélection, on obtient un modèle M_i, disons. Si le modèle M_i est adéquat et qu'on a plusieurs milliers d'essais, alors $D(\widehat{\boldsymbol{\beta}}_{M_i}) \dot{\sim} \chi^2_{JK-p_i}$ et son espérance devrait être approximativement égale à $JK - p_i$. Modélisez les données `cancer` avec un modèle de régression logistique (loi binomiale et fonction liaison logit). La base de données contient deux variables explicatives catégorielles, `age` et `maligne`, qui ont trois et deux niveaux respectivement. Faites une analyse de déviance et sélectionnez le meilleur modèle par élimination en partant du modèle saturé (procédure descendante).

Solution

Il y a $JK = 6$ paramètres dans le modèle saturé.

modèle	déviance	p	covariables
M_0	12,66	1	`aucune`
M_1	6,64	3	`age`
M_2	5,96	2	`maligne`
M_3	0,49	4	`age`, `maligne`

TABLE 1 – Analyse de déviance pour les données `cancer`

On compare d'abord le modèle additif, M_3, au modèle saturé avec $JK = 6$ coefficients. La déviance est de 0,49 et, sous l'hypothèse nulle $\mathscr{H}_0 : \nu_{jk} = 0, j = 1, \ldots, J-1, k = 1, \ldots, K-1$, $D(\widehat{\boldsymbol{\beta}}_{M_3}) = 0,49 \sim \chi^2_2$ asymptotiquement. La valeur-p du test est de 0,78, donc on ne rejette pas l'hypothèse nulle que le modèle additif M_3 est une simplification adéquate du modèle saturé.

On peut ensuite comparer le modèle M_3 aux modèles M_2 et M_1. Considérons d'abord M_3 versus M_2, ce qui correspond à l'hypothèse nulle $\mathscr{H}_0 : \beta_j = 0 (j = 1, \ldots, J-1)$; la statistique du test du rapport de vraisemblance $D(\widehat{\boldsymbol{\beta}}_{M_2}) -$

$D(\widehat{\boldsymbol{\beta}}_{M_3}) \dot{\sim} \chi^2_{p_3-p_2}$; on compare la valeur de la statistique, 5,46, aux quantiles de la loi χ^2_2. Le 95% percentile de la loi χ^2_2 est 5,99, donc on ne rejette pas l'hypothèse nulle à niveau 5% que M_2 est une simplification adéquate de M_3. La valeur-p est de $0,065$.

On aurait aussi pû comparer le modèle M_3 au modèle M_1 ; la statistique du test de rapport de vraisemblance est $6.64 - 0.49 = 6.15$; cette valeur est à comparer au 95% de la loi χ^2_1, soit 3,84. La valeur-p est 0,013, donc le coefficient associé à `maligne` est significativement différent de zéro.

Finalement, on peut comparer M_2 à M_0. L'hypothèse nulle est $\mathscr{H}_0 : \gamma_k = 0 (k = 1, \ldots, K-1)$ et la valeur-p est 0,009 ; les paramètre pour `maligne` est une fois de plus différent de zéro. ($\star$) Si le modèle M_2 était adéquat, la déviance suivrait approximativement une loi χ^2_4 (mais la loi de référence ici dépend des vrais paramètres du modèle, lesquels sont inconnus). La déviance de M_2 est 5,96 et la valeur-p apppproximative est 0,25. Cette approximation pourrait être douteuse dans la mesure où le nombre pour chaque scénario du plan d'expérience est fortement débalancé (il y a moins de patients dans les tranches d'âge supérieures). Une meilleure façon de vérifier l'adéquation est de simuler des nouvelles données du modèle ajusté en conditionnant sur le nombre de patients par tranche d'âge et le status du cancer pour obtenir une loi empirique pour la déviance et la comparer à la loi χ^2_4 ; dans notre exemple, il semble que cette dernière soit adéquate (voir code R).

4.6 **Équivalence entre les modèles de régression Poisson et binomiale :** Si $Y_j \sim \mathsf{Bin}(m_j, p_j)$ et $m_j p_j \to \mu_j$ quand $m_j \to \infty$, on peut approximer la loi de Y_j par une loi Poisson $\mathsf{Po}(\mu_j)$. De ce fait, on pourrait considérer un modèle linéaire généralisé avec

$$\log(\mu_j) = \log(m_j) + \log(p_j).$$

Dans ce modèle, m_j est une constante fixe et le coefficient du prédicteur $\log(m_j)$ est exactement un ; c'est un terme de décalage.

(a) Ajustez les modèles $M_0, \ldots, M_3$ à l'aide d'une vraisemblance binomiale et une fonction de liaison logistique pour les données `fumeurs`. Ce jeu de données contient le nombre de cas de cancer du poumon par tranche d'âge (`age`) et par habitude (`fume`). Rapportez la valeur de la déviance et le nombre de degrés de liberté du modèle (nombre de paramètres). Faites une analyse de déviance séquentielle descendante. Répétez votre analyse avec une régression de Poisson possédant un terme de décalage.

(b) Comparez les résultats obtenus à l'aide du modèle de régression logistique et de régression de Poisson en terme de probabilité de décès pour chaque catégorie.

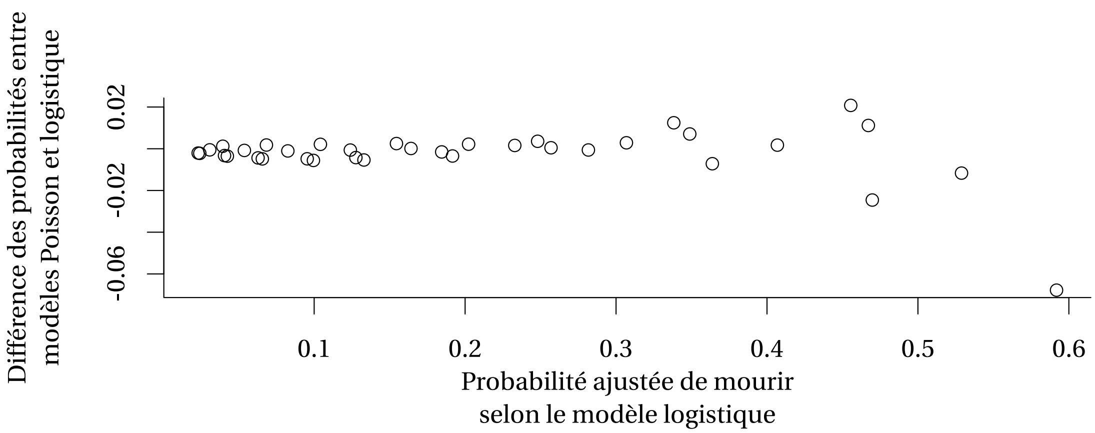
Solution

(a) On procède par élimination séquentielle descendante en partant du modèle saturé avec $m = 36$ paramètres, qu'on compare d'abord au modèle M_3 ; cette comparaison revient à tester $\mathscr{H}_0 : \nu_{jk} = 0, j = 1, \ldots, J-1; k = 1, \ldots, K-1$. La valeur-$p$ est 0,61 pour le modèle de Poisson et 0,55 pour le modèle binomial. On n'a pas de preuve que le modèle M_3 n'est pas une simplification adéquate du modèle saturé. Le modèle saturé ne peut pas être davantage simplifié et semble adéquat dans la mesure où $D(\widehat{\boldsymbol{\beta}}_{M_3}) \approx m - p = 24$.

modèle	déviance (binom.)	déviance (Poisson)	p
M_0	4055,98	4917,03	1
M_1	3910,70	4740,34	4
M_2	191,72	247,94	9
M_3	21,49	22,44	12

TABLE 2 – Analyse de déviance pour les données `fumeurs`

(b) On voit que, pour certaines catégories pour lesquelles le taux de mortalité relié au cancer du poumon est élevé, l'approximation de Poisson à la loi binomiale n'est pas bonne et les différences sont plus marquées entre les deux modèles.


Différence des probabilités entre
modèles Poisson et logistique
0.02
-0.02
-0.06
0.1
0.2
0.3
0.4
0.5
0.6
Probabilité ajustée de mourir
selon le modèle logistique