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.

(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.

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.

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.
- 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 ?

(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 ?

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

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

(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 li-

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

néaire 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.*

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`.
- 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.

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

(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.

(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.

(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?

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

tandis que les moyennes ajustées sont continues.

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 ∞,

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

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).

4.6 **Équivalence entre les modèles de régression Poisson et binomiale :** Si $Y_j \sim \text{Bin}(m_j, p_j)$ 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 $\text{Po}(\mu_j)$. 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.

Compilé le 17/11/2020 à 11:15