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 4.4.1 : modèle de régression de Poisson avec la covariable `fds`.
- modèle 4.4.2 : modèle de régression de Poisson, en incluant les covariables `fds`, `humid` et `temp`.
- modèle 4.4.3 : modèle de régression binomiale négative incluant les covariables `fds`, `humid` et `temp`.
- modèle 4.4.4 : modèle de régression binomiale négative incluant les variables explicatives `jour` (catégorielle), `humid` et `temp`.

(a) Est-ce que le modèle 4.4.1 représente une simplification adéquate du modèle 4.4.2? 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 4.4.2.

(c) Quel modèle choisiriez-vous parmi les deux modèles 4.4.2 et 4.4.3? 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 4.4.4 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 4.4.4 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 4.4.3 au modèle 4.4.4 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 **Parties de soccer** : Soit Y_{ij} (resp. Z_{ij}) le score de l'équipe à domicile (resp. visiteurs) pour une partie de soccer

tandis que les moyennes ajustées sont continues.

opposant les équipes i et j. Maher (1982) a suggéré de modéliser les scores via

$$Y_{ij1} \sim \text{Po}\{\exp(\delta + \alpha_i + \beta_j)\}, \quad Y_{ij2} \sim \text{Po}\{\exp(\alpha_j + \beta_i)\}, \qquad i \neq j; i, j \in \{1, \ldots, 20\}, \tag{E4.5.1}$$

où α_i représente la puissance offensive de l'équipe i, β_j la puissance défensive de l'équipe j et δ l'avantage du terrain commun à toutes les équipes qui jouent à domicile. Les points des deux équipes et de parties différentes sont supposés indépendants les uns des autres. La base de donnée `soccer` contient les résultats des parties de soccer pour la saison 2015 de la *English Premier Ligue* (EPL), avec les variables suivantes :

— `buts` : nombre de buts comptés durant la partie par `equipe`.
— `equipe` : variable catégorielle indiquant le nom de l'équipe qui a compté les buts.
— `adversaire` : variable catégorielle donnant le nom de l'adversaire.
— `domicile` : variable binaire, 1 si `equipe` joue à domicile, 0 sinon.

Note : les buts d'une même partie ne sont pas adjacents parce que la base de données est triée par domicile. Par exemple, la première partie oppose Manchester U. (1 but, ligne 1) à domicile à Tottenham (0 but, ligne 381).

(a) Un avantage du terrain δ commun à toutes les équipes est logique pourvu que le nombre de buts comptés à domicile soient indépendants de ceux comptés à l'extérieur, c'est-à-dire en l'absence de terme d'interaction entre les pointages. Pour vérifier cette hypothèse, on agrège les pointages de chaque partie durant la saison et on construit un tableau de contingence à deux facteurs (Table 1) pour les points de l'équipe à domicile versus ceux de l'équipe adverse. Le fichier `socceragg` contient ce tableau de contingence en format long ; à l'aide de ces données, testez l'hypothèse d'indépendance et concluez.

	visiteur						
domicile	0	1	2	3	4	5	6
0	32	33	9	14	3	0	1
1	37	41	28	11	3	1	0
2	27	25	29	7	2	1	0
3	18	15	10	5	2	0	0
4	9	6	3	0	0	1	0
5	0	4	0	0	0	0	0
6	0	1	2	0	0	0	0

TABLE 1 – Nombre d'occurrences de scores pour les parties de soccer de l'EPL

(b) Ajustez le modèle de l'Equation (E4.5.1) et répondez aux questions suivantes :
   i. À l'aide du modèle ajusté, donnez le nombre moyen de buts comptés par chaque équipe dans une partie opposant Manchester United (à domicile) et Liverpool.
   ii. Rapportez et interprétez l'avantage du terrain estimé, $\widehat{\delta}$.
   iii. Testez si l'avantage du terrain δ est significativement différent de zéro.
   iv. La loi asymptotique de la statistique de déviance D est χ^2_{n-p-1}. Cette approximation n'est valable que quand le nombre d'observations dans chaque groupe est grand, or seules 38 parties sont disputées par chaque équipe à domicile et à l'extérieur. On peut approximer plutôt la loi nulle de D par simulation en répétant $B = 10\,000$ fois les étapes suivantes :
      A. générer de nouvelles données de loi Poisson du modèle ajusté,
      B. ajuster la régression de Poisson donnée par l'Equation (E4.5.1) aux données simulées,
      C. calculer la statistique de déviance.

   Le Table 2 donne les quantiles empiriques des $B = 10\,000$ valeurs de déviance obtenues sur les données simulées. Commentez sur la qualité de l'ajustement sur la base de la statistique de déviance et du Table 2. Contrastez vos conclusions avec celles obtenues en utilisant l'approximation asymptotique χ^2 de la loi

niveau	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
quantile	760.53	770.33	782.30	803.41	826.25	849.92	871.23	885.74	897.44

TABLE 2 – Quantiles des statistiques de déviance simulées selon le modèle de Maher (1982)

nulle de la déviance.

(c) Maher a aussi suggéré des modèles plus complexes, dont un dans lequel la force offensive et défensive différait selon que l'équipe soit à domicile ou à l'extérieur, à savoir

$$Y_{ij1} \sim \text{Po}\{\exp(\alpha_i + \beta_j + \delta)\}, \quad Y_{ij2} \sim \text{Po}\{\exp(\gamma_j + \omega_i)\}, \qquad i \neq j;\ i, j \in \{1, \ldots, 20\} \tag{E4.5.2}$$

Est-ce que le modèle (E4.5.2) est significativement meilleur que le modèle (E4.5.1) ?

(d) Est-ce qu'un modèle de Poisson similaire serait adéquat pour le pointage d'une partie de basketball ? Expliquez votre réponse. *Indice : quel est le nombre de points comptés en moyenne lors d'une partie ?*

4.6 **Bush vs Gore** : L'élection présidentielle américaine de 2000 opposait Georges W. Bush (Républicain) et Albert A. Gore (Démocrate), ainsi qu'une quantité de candidats marginaux. La Floride a été remporté par Bush par une mince marge de 537 votes, ce qui en a fait l'état pivot déterminant le vainqueur de l'élection. Plusieurs analystes ont affirmé que les bulletins de vote de type papillon utilisés dans les quartiers pauvres du comté de Palm Beach ont mené à la confusion d'électeurs et ont dépossédé Al Gore de quelques milliers de voix qui ont été à un tiers candidat d'extrême droite, Patrick Buchanan (Réforme). Smith (2002) a analysé les résultats de l'élection par comté et s'est intéressé au décompte du comté de Palm Beach, dans lequel un nombre anormalement élevé d'électeurs (3407) exprimaient une préférence pour Buchanan.

Le but de cet exercice est de construire un modèle pour prédire le nombre moyen de votes pour Buchanan dans le comté de Palm Beach sur la base de ses résultats dans les autres comtés. Les données `buchanan` contiennent les variables suivantes pour chacun des 67 comtés de l'État de Florida :

- `comte` : nom du comté
- `popn` : population du comté en 1997.
- `blanc` : pourcentage de « blancs » [*sic*] en 1996 (selon le questionnaire du recensement ; personnes ayant des origines européennes, nord-africaines ou moyen-orientales (définition du recensement).
- `noir` : pourcentage de « noirs » [*sic*] ou Afro-Américains (personnes ayant des origines d'Afrique sub-saharienne), en 1996.
- `hisp` : pourcentage d'hispaniques (groupe ethnique) en 1996.
- `a65` : pourcentage de la population âgée de plus de 65 ans selon les estimés de population de 1996 et 1997.
- `dsec` : pourcentage de la population avec un diplôme d'études secondaires (recensement de 1990).
- `coll` : pourcentage de la population diplômée d'un collège (recensement de 1990).
- `revenu` : revenu moyen individuel en 1994.
- `buch` : nombre de votes pour Pat Buchanan.
- `bush` : nombre de votes pour Georges W. Bush.
- `gore` : nombre de votes pour Al Gore.
- `totmb` : nombre total de bulletins de vote, moins les votes pour Buchanan.

(a) Calculez la proportion totale de votes obtenus pour Buchanan pour toute la Floride.

(b) Produisez un graphique du pourcentage de votes obtenus par Buchanan, `buch`/(`buch`+`totmb`), versus ln(`popn`) et commentez.

Excluez les résultats de Palm Beach pour le reste de la question en remplaçant la valeur de `buch` par une valeur manquante.

(c) On considère d'abord un modèle de Poisson pour le pourcentage de votes obtenus par Buchanan, `buch`/`totmb`,

en fonction de `blanc`, ln(`hisp`), `a65`, `dsec`, ln(`coll`), `revenu`.

i. Expliquez l'utilité d'utiliser un terme de décalage dans ce problème.

ii. Pourquoi est-ce que `totmb` est un meilleur choix de dénominateur que `popn` pour le taux d'appuis? Expliquez.

iii. Est-ce que le modèle de régression de Poisson est approprié? Justifiez votre réponse.

iv. Expliquez pourquoi, s'il y a des preuves de surdispersion, cela implique forcément qu'un modèle de régression binomiale est inadéquat. *Indice : quelle est la variance de la loi binomiale et comment cette dernière se compare à la loi Poisson pour un taux?*

(d) Ajustez le même modèle, mais cette fois-ci avec une loi binomiale négative. Utilisez ce modèle pour prédire le nombre moyen de votes pour Buchanan dans le comté de Palm Beach. Commentez sur la différence entre le nombre de votes obtenus par Buchanan et cette prédiction.

4.7 **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).

4.8 **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 15/12/2020 à 22:15