Instructions :

— Répondez aux questions suivantes à l'aide de SAS et fournissez le code utilisé pour vos analyses dans un fichier (extension `.txt`, encodage `utf8`).
— Votre rapport doit être remis en ligne en format PDF et en version papier en classe et ne devrait pas faire plus de 15 pages ; toute page excédentaire sera ignorée lors de la correction. Soyez concis mais précis : n'incluez que les sorties pertinentes.
— Les erreurs sont pénalisées même si elles ne sont pas en lien direct avec la question.
— Des exemples de prédiction de modèles linéaires généralisés (procédure `genmod`) sont disponibles sur Piazza.

1.5 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

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

où l'espérance $\lambda_i = \exp(\beta_0 + \beta_1 \texttt{solde}_i)$ et $\texttt{solde}_i$ 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.*
(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.
(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%.
(d) Votre gérant veut savoir si le profits quotidiens moyens pendant les soldes sont différentes 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**).

2.1 **Parties de soccer** : Soit Y_{ij} (resp. Z_{ij}) le score de l'équipe à domicile (resp. visiteurs) pour une partie de soccer opposant les équipes i et j. Maher (1982) a suggéré de modéliser les scores via

$$Y_{ij} \sim \mathsf{Po}\{\exp(\delta + \alpha_i + \beta_j)\}, \quad Z_{ij} \sim \mathsf{Po}\{\exp(\alpha_j + \beta_i)\}, \qquad i \neq j; i, j \in \{1, \ldots, 24\}, \tag{E2.2}$$

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.
— `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.

(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 (Tableau 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.

Compilé le 26/10/2020 à 20:16

	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

Tableau 1 – Nombre d'occurences de scores pour les parties de soccer de l'EPL

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

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

(b) Ajustez le modèle de l'Équation (E2.2) 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}, 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
      A. générant de nouvelles données de loi Poisson du modèle ajusté,
      B. ajustant la régression de Poisson donnée par l'Équation (E2.2) aux données simulées,
      C. calculant la statistique de déviance.

      Le Tableau 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 de Tableau 2. Contrastez vos conclusions avec celles obtenues si on utilisant l'approximation asymptotique χ^2 de la loi nulle.

(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

$$Y_{ij} \sim \mathsf{Po}\{\exp(\alpha_i + \beta_j + \delta)\}, \quad Z_{ij} \sim \mathsf{Po}\{\exp(\gamma_j + \omega_i)\}, \qquad i \neq j;\ i, j \in \{1, \ldots, 24\} \tag{E2.3}$$

Est-ce que le modèle (E2.3) est significativement meilleur que le modèle (E2.2)?

(d) Est-ce qu'un modèle de Poisson similaire serait adéquat pour le pointage d'une partie de basketball? Expliquez votre réponse.

2.2 **Bush vs Gore** : L'élection présidentielle américaine de 2000 opposait Georges W. Bush (Républicain) et Albert A. Gore (Démocrate). La Floride et ses 25 grands électeurs 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 utilisé dans les quartiers pauvres ont mené à la confusion d'électeurs et ont dépossédé Al Gore de quelques milliers de voix qui ont été à 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. Les données `floride` contiennent les

variables suivantes pour chacun des 67 comtés de l'État :

- `comte` : nom du comté
- `popn` : population du comté en 1997.
- `blanc` : pourcentage de blancs en 1996.
- `noir` : pourcentage de noirs en 1996.
- `hisp` : pourcentage d'hispaniques 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).
- `revenue` : 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.

Note : il y a un chevauchement entre `blanc`/`noir` et `hisp`, de telle sorte que les trois catégories additionnées ne donnent pas 100%. Puisque nous sommes intéressés à prédire le nombre de votes de Buchanan, on exclut ces derniers du nombre total de bulletins déposés.

(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.

On cherche à construire un modèle pour prédire le nombre moyen de votes pour Buchanan dans le comté de Palm Beach (comté 50), en utilisant uniquement l'information des autres comtés. **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 nombre de votes pour Buchanan, `buch`, en fonction de `blanc`, ln(`hisp`), `a65`, `dsec`, ln(`coll`), `revenue`, avec `totmb` comme terme de décalage.

   i. Expliquez l'utilité du terme de décalage dans ce problème.

   ii. Pourquoi est-ce que `totmb` est un meilleur choix pour le terme de décalage que `popn`? Expliquez.

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

(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.

(e) 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?*

Compilé le 26/10/2020 à 20:16