

**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

$$P(Y_i = y_i | x_i) = \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, \dots$$

où l'espérance  $\lambda_i = \exp(\beta_0 + \beta_1 \text{solde}_i)$  et  $\text{solde}_i$  est un indicateur binaire qui vaut un lors des soldes et zéro autrement.

- Dérivez l'estimateur du maximum de vraisemblance pour  $(\beta_0, \beta_1)$ . *Indice : les estimateurs du maximum de vraisemblance sont invariants aux reparamétrisations. Considérez les deux échantillons solde/hors solde.*
  - 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.
  - Calculez la matrice d'information observée et utilisez-la pour dériver les erreurs-type de  $\hat{\beta}_0$  et de  $\hat{\beta}_1$  et un intervalle de confiance à niveau 95%.
  - 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  $\beta_0$  et  $\beta_1$  (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

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

où  $\alpha_i$  représente la puissance offensive de l'équipe  $i$ ,  $\beta_j$  la puissance défensive de l'équipe  $j$  et  $\delta$  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 :

- **but** : 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.
- Un avantage du terrain  $\delta$  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.

domicile	visiteur						
	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’occurrences 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 :

- 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.
- Rapportez et interprétez l’avantage du terrain estimé,  $\hat{\delta}$ .
- Testez si l’avantage du terrain  $\delta$  est significativement différent de zéro.
- La loi asymptotique de la statistique de déviance  $D$  est  $\chi^2_{n-p-1}$ . 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 :
  - générer de nouvelles données de loi Poisson du modèle ajusté,
  - ajuster la régression de Poisson donnée par l’Équation (E2.2) aux données simulées,
  - calculer 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 du Tableau 2. Contrastez vos conclusions avec celles obtenues en utilisant l’approximation asymptotique  $\chi^2$  de la loi nulle de la déviance.

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

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

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. *Indice : quel est le nombre de points comptés en moyenne lors d’une partie?*

**2.2 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 floride 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,  $\text{buch}/(\text{buch}+\text{totmb})$ , versus  $\ln(\text{popn})$  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,  $\text{buch}/\text{totmb}$ , en fonction de blanc,  $\ln(\text{hisp})$ , a65, dsec,  $\ln(\text{coll})$ , 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.