Instructions:

- Answer the following questions using SAS and provide the code you used to perform the analysis in a separate file (`.txt` extension, `utf8` encoding).
- Your report must be submitted as a PDF file and should not exceed 15 pages; any additional page will be ignored. Be brief, but precise; only include relevant output.
- Errors are penalized even if they are not directly related to the question.
- Sample code is provided on Piazza for predicting from `genmod` procedure output.

1.5 We consider a simple Poisson model for the number of daily sales in a store, which are assumed independent from one another. Your manager tells you the latter depends on whether the store is holding sales or not. The mass function of the Poisson distribution is

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

and we model $\lambda_i = \exp(\beta_0 + \beta_1 \texttt{sales}_i)$, where $\texttt{sales}_i$ is a binary indicator equal to unity during sales and zero otherwise.

(a) Derive the maximum likelihood estimator of (β_0, β_1). *Hint: maximum likelihood estimators are invariant to reparametrization.*

(b) Calculate the maximum likelihood estimates for a sample of size 12, where the number of transactions outside sales is $\{2;5;9;3;6;7;11\}$, and during sales, $\{12;9;10;9;7\}$.

(c) Calculate the observed information matrix and use the latter to derive standard errors for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and a 95% confidence interval for the parameters.

(d) Your manager wants to know if the daily profits during sales period are different from those outside of the sales period. She calculates that the average profit during sales is \$20 per transaction, compared to \$25 normally. Test this hypothesis using a likelihood ratio test. *Hint: write the null hypothesis of equal profit in terms of the model parameters β_0 and β_1.* (**difficult**)

2.1 **Soccer matches**: Let Y_{ij} (resp. Z_{ij}) denote the score of the home (resp. visitor) team for a soccer match opposing teams i and j. Maher (1982) suggested modelling the scores as

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

where α_i represent the offensive strength of the team, β_j the defensive strength of team j and δ is the common home advantage. The scores in a match and between matches are assumed to be independent of one another. The data set `soccer` contains the results of football (soccer) matches for the 2015 season of the English Premier Ligue (EPL) and contains the following variables

- `score`: number of goals during
- `team`: categorical variable giving the name of the team which scored the goals
- `opponent`: categorical variable giving the name of the adversary
- `home`: binary variable, 1 if `team` is playing at home, 0 otherwise.

(a) A common home advantage δ makes sense provided that the scores at home and away are independent, i.e., there is no interaction between the two. To validate this hypothesis, we consider aggregates over multiple seasons of the scores, cross-classified in terms of number of points for the team at home and the team away, for each match (Table 1). The file `socceragg` contains the two-way contingency table in long-format. Using the latter, test the assumption of independence.

(b) Fit the model characterized by Equation (E2.2) and answer the following questions:

i. Using the fitted model, give the expected number of goals for a match between Manchester United (at

Compiled 27/10/2020 at 10:50

	away						
home	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: Frequency of scores for EPL soccer matches in 2015

level	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 of the simulated deviance statistics based on the model of Maher (1982)

home) against Liverpool.

ii. Report and interpret the estimated home advantage $\widehat{\delta}$.

iii. Test whether the home advantage δ is significantly different from zero.

iv. The asymptotic null distribution of the deviance statistic D is χ^2_{n-p-1}, but the latter is only valid when the number of observations in each group is large. In our analysis, there are only 38 matches in a given year at home/visiting for each team. We can instead approximate the null distribution of D using a simulation: specifically, we repeat the following steps $B = 10\,000$ times

   A. generate new Poisson data from the fitted model
   B. fit the Poisson regression specified by Equation (E2.2) on the simulated data
   C. calculate the deviance statistic.

   Table 2 gives quantiles of the simulated null distribution of the deviance based on these $B = 10\,000$ simulations. Comment on the adequacy of the fit based on the deviance statistic and Table 2 and contrast with the conclusions obtained by using the asymptotic null distribution of the deviance.

(c) Maher also suggested more complex models, including one in which the offensive and defensive strength of each team changed depending on whether they were at home or visiting another team, i.e.

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

Does Model (E2.3) fit significantly better than Model (E2.2)?

(d) Why would a similar Poisson wouldn't be adequate to model basketball scores? Explain. *Hint: what is the average score in a basketball match?*

2.2 **Bush vs Gore**: the 2000 US presidential election opposed Georges W. Bush (GOP) and Albert A. Gore (Democrat), as well as marginal third party candidates. The tipping state was Florida, worth 25 electors, which Bush won by a narrow margin of 537 votes. There have been many claims that the design of so-called butterfly ballots used in poor neighborhoods of Palm Beach county led to confusion among voters and that this deprived Gore of some thousands of votes that were instead assigned to a paleoconservative third-party candidate, Patrick Buchanan (Reform). Smith (2002) analysed the election results in Palm Beach country, in which a unusually high number of ballots (3407) were cast for Buchanan.

We are interested in building a model to predict the expected number of votes for Buchanan in Palm Beach county, based only on the information from other county votes. The `buchanan` data contains the following variables:

- `county`: name of county
- `popn`: population of the county in 1997.
- `white`: percentage of white (*sic*) in 1996 (per US Census definitions, people having origins in any of the original peoples of Europe, the Middle East, or North Africa).
- `black`: percentage of Black and African Americans in 1996 (origins from sub-saharian Africa).
- `hisp`: percentage of Hispanics in 1996.
- `geq65`: percentage of the population aged 65 and above based on 1996 and 1997 population estimates.
- `highsc`: percentage of the population with a high school degree (1990 Census data).
- `coll`: percentage of the population that are college graduates (1990 Census data).
- `income`: mean personal income in 1994.
- `buch`: total ballots cast for Pat Buchanan (Reform).
- `bush`: total ballots cast for Georges W. Bush (GOP).
- `gore`: total ballots cast for Al Gore (Democrat).
- `totmb`: total number of votes cast for the presidential election in each county, minus Buchanan votes.

(a) Calculate the total proportion of votes for Buchanan in Florida.

(b) Plot the percentage of votes obtained by Buchanan (`buch/(buch+totmb)`) against ln(`popn`) and comment.

Exclude the results of Palm Beach county for the rest of the question.

(c) We consider first a Poisson model for the percentage of votes for Buchanan, `buch/totmb`, as a function of `white`, ln(`hisp`), `geq65`, `highsc`, ln(`coll`), `income`.

i. Explain why an offset is necessary in this case.

ii. Why is `totmb` a better choice of denominator than `popn` for the rate? Explain.

iii. Is the Poisson model appropriate? Justify your answer.

iv. Explain why, if there is evidence of overdispersion, this means the binomial model is also inadequate. *Hint: what is the variance of the binomial distribution and how does it relate to the Poisson distribution?*

(d) Use a negative binomial model with the same covariates to predict the expected number of Buchanan votes in Palm Beach county. Comment hence on the discrepancy between this forecast and the number of votes received in the election.

Compiled 27/10/2020 at 10:50