# MATH 60604
# Modélisation statistique
# § 4g - Taux et termes de décalage

Léo Belzile

HEC Montréal
Département de sciences de la décision

# Décalage et comparaisons de dénombrement

- Jusqu'à présent, nous avons supposé que la variable de dénombrement $Y$ était comparable d'un individu à l'autre.
  - Dans l'exemple d'achats, $Y_i$ représentait le nombre de fois que le sujet $i$ avait acheté le produit dans le mois suivant l'étude.
- Et si la période de suivi variait d'un individu l'autre?
  - le nombre d'accidents de travail dans une entreprise pour une période donnée dépend du nombre d'employés.
  - le nombre de cancer par région dépend du nombre d'habitants.

Si les nombres ne sont pas comparables, on peut considérer les **taux** (nombre d'achats par mois, nombre d'accident par employé, etc.)
Si on modélise le taux avec le modèle de Poisson, ce dernier est adéquat **seulement** si le taux est faible.

# Données sur les accidents de la route

La *National Highway Traffic Safety Administration* (NHTSA) compile des statistiques sur le nombre de morts sur les routes aux États-Unis. Les données `accident` dénombre les décès en 2010 et en 2018 par États, recensés par régions géographiques (`region`) telles que définies par la NHTSA et catégorisées selon le `moment` où l'accident a eu lieu (jour ou nuit).

- Soit $Y_i$ le nombre de décès à un moment donné durant une année donnée pour la région $i$;
- Soit $N_i$ le nombre d'habitants dans la région $i$.

Notre objectif est d'estimer la relation entre nombre d'accident fatal selon le moment de la journée et l'année.

**Fatal Motor Vehicle Crashes**[1]
**Note: Click on the link within a table cell to map crash locations**

| Crash Date (Year) by NHTSA Region | | Time Of Day | | | |
|---|---|---|---|---|---|
| | | Daytime | Nighttime | Unknown | Total |
| 2010 | 1 = ME, MA, NH, RI, VT | 196 | 210 | 1 | 407 |
| | 2 = CT, NJ, NY, PA, PR | 917 | 964 | 1 | 1,882 |
| | 3 = DE, DC, KY, MD, NC, VA, WV | 539 | 642 | 2 | 1,183 |
| | 4 = AL, FL, GA, SC, TN | 1,233 | 1,613 | 13 | 2,859 |
| | 5 = IL, IN, MI, MN, OH, WI | 899 | 1,005 | 1 | 1,905 |
| | 6 = LA, MS, NM, OK, TX | 810 | 1,307 | 7 | 2,124 |
| | 7 = AR, IA, KS, NE, MO | 295 | 313 | 1 | 609 |
| | 8 = CO, NV, ND, SD, WY, UT | 240 | 241 | 0 | 481 |
| | 9 = AZ, CA, HI | 791 | 1,100 | 20 | 1,911 |
| | 10 = AK, ID, MT, OR, WA | 140 | 211 | 1 | 352 |
| | Total | 6,060 | 7,606 | 47 | 13,713 |

# Accidents de la route et décalage

- Si on ignore la taille de la population, le modèle de régression de Poisson (ou binomiale négative) s'écrirait

$$\ln(\mu_i) = \ln\{\mathsf{E}\,(Y_i)\} = \beta_0 + \beta_1 \texttt{moment} + \beta_2 \texttt{annee}$$

- Si on prend en compte la taille de la population, cela revient à modéliser le taux $Y_i/N_i$ plutôt que $Y_i$.
- On fixe

$$\ln\left\{\frac{\mathsf{E}\,(Y_i)}{N_i}\right\} = \beta_0 + \beta_1 \texttt{moment} + \beta_2 \texttt{annee}$$

ou de manière équivalente

$$\ln\{\mathsf{E}\,(Y_i)\} = \beta_0 + \beta_1 \texttt{moment} + \beta_2 \texttt{annee} + \ln(N_i)$$

- Le terme $\ln(N_i)$ est un **terme de décalage**; une variable explicative incluse sans paramètre.

# Régression binomiale négative pour `accidents`

## Code SAS pour inclure un terme de décalage

```
data accident;
set modstat.accident;
logpopn=log(popn);
run;
proc genmod data=accident;
class moment(ref="jour") annee(ref="2010");
model nmorts=moment annee / dist=negbin link=log
    offset=logpopn type3 lrci;
run;
```

L'option `offset` pourrait aussi être utilisée dans une régression de Poisson.

| Informations sur le modèle | |
|---|---|
| **Table** | WORK.ACCIDENT |
| **Distribution** | Negative Binomial |
| **Fonction Link** | Log |
| **Variable dépendante** | nmorts |
| **Variable de décalage** | logpopn |

# Interprétation des paramètres avec décalage

**Analyse des paramètres estimés du maximum de vraisemblance**

| Paramètre | | DDL | Estimation | Erreur type | Rapport de vraisemblance Intervalle de confiance à95% | | Khi-2 de Wald | Pr > khi-2 |
|---|---|---|---|---|---|---|---|---|
| **Intercept** | | 1 | -10.9062 | 0.0706 | -11.0456 | -10.7622 | 23869.7 | <.0001 |
| **moment** | nuit | 1 | 0.2266 | 0.0816 | 0.0628 | 0.3903 | 7.72 | 0.0055 |
| **moment** | jour | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **annee** | 2018 | 1 | 0.2300 | 0.0816 | 0.0662 | 0.3938 | 7.95 | 0.0048 |
| **annee** | 2010 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **Dispersion** | | 1 | 0.0648 | 0.0147 | 0.0426 | 0.1043 | | |

- La variable de décalage $\ln(N)$ n'apparaît pas dans le tableau.
- La statistique de déviance (sortie omise) est $40{,}269$ pour $37$ degrés de liberté (rapport de $1{,}09$). La valeur-$p$ correspondante est $0{,}327$, donc il n'y a pas de preuve que notre modèle est inadéquat.
- Le taux de mortalité durant le jour en 2010 est $\exp(\widehat{\beta}_0) = \exp(-10{,}91)$ ou $1{,}83/100000$, soit un taux de $1{,}83$ décès par $100\,000$ habitants (avec intervalle de confiance à 95% $[1{,}60, 2{,}12] \times 10^{-5}$).
- On estime que la mortalité moyenne entre 2010 et 2018 augmente de 26%, puisque $\exp(0{,}23) = 1{,}26$.