

Travail de session 2/2

Remis

Le vendredi 17 avril 2018

Dans le cadre du cours

«Probabilités et statistiques»

MTH2302D

Par Charles-Édouard Turcot

Matricule 1816907

Groupe 02

École Polytechnique de Montréal



PARTIE 1 : REPRISE DU RAPPORT 1

Contexte général des données

Dans le cadre de la fusillade récente à Parkland, en Floride, un grand débat sur le contrôle des armes à feu a jailli sur la scène politique américaine. Un côté met le blâme de cette fusillade sur l'accessibilité trop facile aux armes à feu tandis que l'autre voudrait que les étudiants et les enseignants dans les écoles soient armés en tout temps. Dans ce contexte, il est intéressant d'observer les données liées à ce sujet à travers le monde. En effet, le contrôle des armes à feu est un sujet controversé et de nombreuses différentes approches ont été prises à travers le monde limitant à divers niveaux l'accès aux armes à feu. De telles limites réduisent donc la quantité de fusils en circulation, tandis qu'un manque de contrôle l'augmente. C'est donc pour ces raisons qu'il faut analyser ces données et voir le résultat du contrôle à travers le monde.

Provenance des données

Les données de population proviennent du site web de la Banque mondiale (<https://data.worldbank.org/>). Elles ont été recueillies par les diverses agences de statistiques de pays à travers le monde à l'aide de recensements de population. Le choix des pays a été fait arbitrairement par l'auteur de ce rapport selon leur importance sur la scène mondiale ainsi que sur l'accessibilité et la crédibilité des données. C'est pour cette raison que peu de pays en développement se retrouvent dans cette analyse. Les données concernant les victimes d'armes à feu et le nombre d'armes à feu en possession de civils ont été récoltées sur le site « GunPolicy » (<http://www.gunpolicy.org>). Il s'agit d'un site web, hébergé par l'école de santé publique de l'université de Sidney, regroupant les données disponibles vis-à-vis de la violence due aux armes à feu. Les données disponibles sur le site proviennent d'études par des organismes internationaux indépendants et aussi de sources gouvernementales. La majorité des données de ce site qui seront utilisées proviennent plus précisément du « Small arms survey », un centre d'étude international indépendant situé au sein de l'Institut de hautes études internationales et du développement de Genève (<http://www.smallarmssurvey.org>). Ils font des rapports annuels fournissant des informations sur tous les aspects des armes à feu légères à travers le monde. Ils récoltent leurs informations en réunissant toutes les sources de données disponibles au sujet des armes à feu : données gouvernementales, estimations par les organisations présentes dans les différents pays, sondages menés directement auprès de la population et données d'enregistrement des armes.

Description de la forme des données

Il y a 4 variables qui ont été récoltées pour chacun des 50 pays que nous considérons dans le cadre de ce travail. Tout d'abord, deux données discrètes : la population et l'année de récolte des données. La population de chaque pays est comptée en nombre d'habitants. Pour chaque pays, les données proviennent d'une même année afin de permettre des chiffres cohérents entre eux. L'année de récolte est donc aussi une variable que nous considérerons. Il y a aussi deux variables

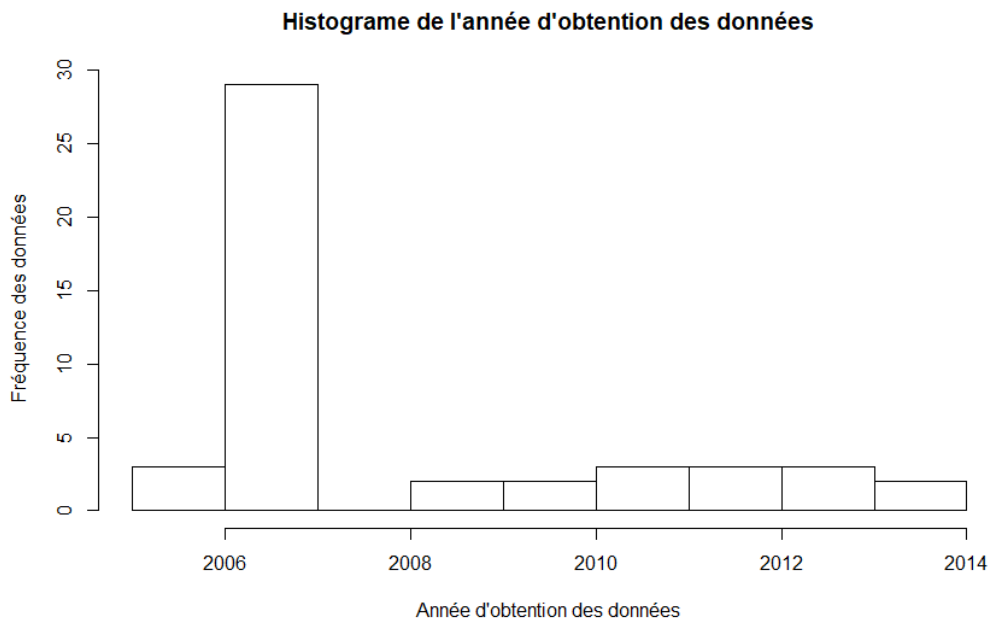
continues : la quantité d'armes à feu et la quantité de morts dues à des armes à feu. La quantité d'armes à feu pour chaque pays est mesurée en nombre d'armes, légales ou non, possédées par des civils dans un pays. Les armes possédées par les représentants de l'état ne sont pas prises en compte. On mesure cette donnée en nombre d'arme par 100 habitants. La quantité de morts dues à des armes à feu comprend les morts accidentelles, suicides et homicides. Cette quantité est mesurée en morts par 100 000 habitants.

Questions Ouvertes

Dans le contexte de ces données, la question principale qu'il faut se poser et qui est à la source du débat aux États-Unis est la suivante : quelle est la corrélation entre l'accessibilité aux armes à feu et le nombre de morts dues à des armes à feu? Dans le cadre de cette analyse, nous considérerons la quantité d'armes à feu par capita comme une mesure directe de l'accessibilité des armes dans un pays. Un autre élément qu'il vaut la peine d'analyser est s'il y a une tendance mondiale à une augmentation ou une baisse du contrôle lors des dernières années. Il faudra donc prendre en compte les années de provenance des données et ensuite les comparer entre elles. La dernière question qu'il vaut la peine de soulever dans ce sujet est à propos de la population. La population d'un pays a-t-elle un effet sur la violence due aux armes à feu au sein de celui-ci? L'analyse des données récoltées permettra de répondre d'une façon simplifiée à ces questions qui ont une importance particulière pour nos voisins du sud en ces temps où la politique américaine est sur nos écrans presque constamment.

PARTIE 2 : ANALYSE DES DONNÉES

Statistique descriptive



Scatter plot showing the relationship between the number of firearms per 100 inhabitants (X-axis) and the number of deaths per 100 inhabitants (Y-axis). The X-axis ranges from 0 to 100, and the Y-axis ranges from 0 to 10. Data points are labeled with country names. Most countries cluster at low values for both variables, with a few outliers showing high mortality rates at low firearm ownership, such as South Africa and Argentina. The United States is a notable outlier with high values for both variables.

[illegible]

Tout d'abord, on affiche les années d'obtention des données par histogramme puisque l'on veut établir s'il y a un lien entre l'année d'obtention des données et la violence due aux armes à feu. Cela à l'avantage de permettre d'avoir rapidement une idée de la répartition des données dans

le temps et de comparer le poids des différentes années. Ensuite, puisqu'on se questionne sur la corrélation entre les morts par armes à feu et le nombre d'armes à feu et entre les morts par armes à feu et la population, on représente ces deux paires de données sur des nuages de points. Cette représentation permet rapidement d'observer s'il y a des chances de corrélation et permet aussi d'identifier facilement les données de chaque pays. Le désavantage de cette représentation est qu'elle ne permet pas de bien évaluer la distribution des données puisque les points denses sur les nuages de points sont difficiles à bien lire.

Régression

Pour les deux paires de données, nous allons tenter de faire une régression linéaire afin de représenter le lien entre les variables.

On obtient d'abord les résultats suivants pour le lien entre la quantité d'armes à feu et la quantité de morts par armes à feu.

```
Call:
lm(formula = mort ~ arm)

Residuals:
    Min       1Q   Median       3Q      Max
-2.4983 -0.9181 -0.4267  0.3834  5.3280

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.90132    0.29942   3.010  0.00427 **
arm          0.07272    0.01306   5.566 1.37e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ce tableau permet donc de dire que $\beta_0 = 0.90132$; $\beta_1 = 0.07272$

Le test de $H_0 : \beta_0 = 0$ contre $H_1 : \beta_0 \neq 0$. On a $t_0 = 0.90132/0.29942 = 3.010$.
valeur-p= 0,00427, donc on rejette H_0 .

Le test de $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$. On a $t_0 = 0.07272/0,01306 = 5.566$.
Puisque valeur-p= 1.37e-06, on rejette H_0 .

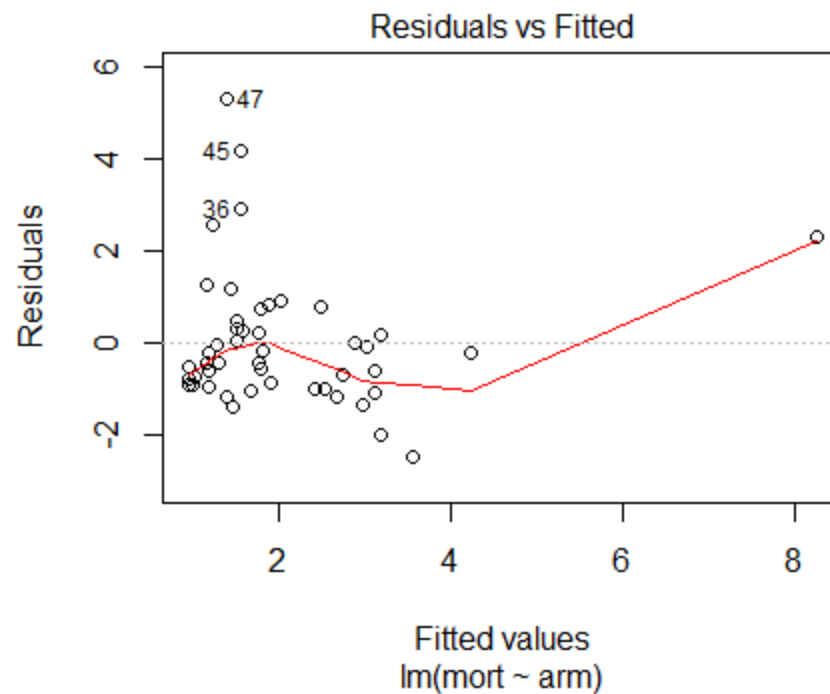
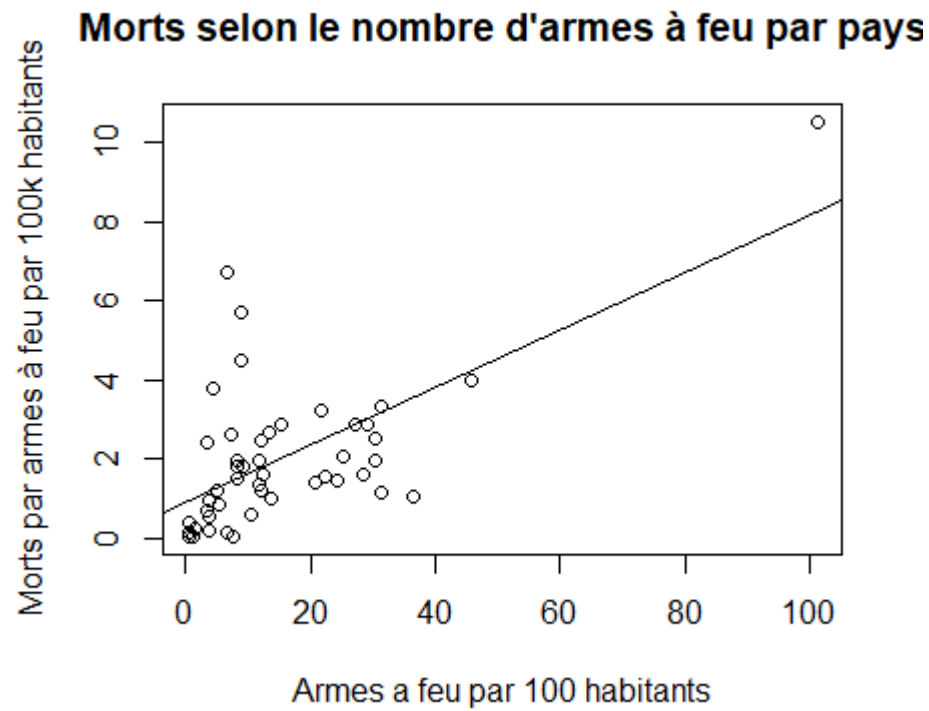
Residual standard error: 1.498 on 45 degrees of freedom
Multiple R-squared: 0.4078, Adjusted R-squared: 0.3946
F-statistic: 30.98 on 1 and 45 DF, p-value: 1.369e-06

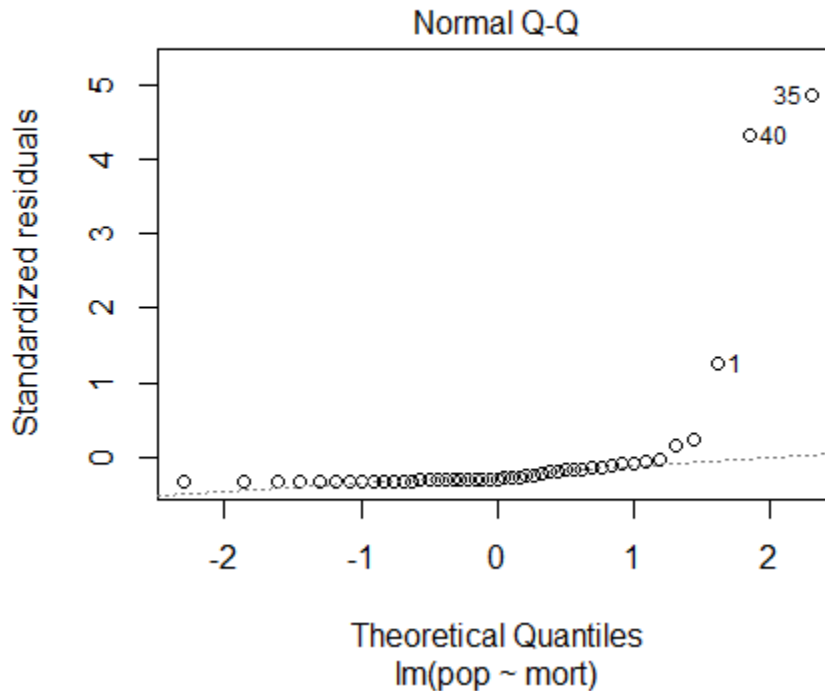
```
> confint(Reglin)
                2.5 %    97.5 %
(Intercept) 0.29825548 1.5043755
arm         0.04640689 0.0990333
```

```
> anova(Reglin)
Analysis of Variance Table
```

```
Response: mort
      Df Sum Sq Mean Sq F value    Pr(>F)
arm     1  69.542   69.542   30.983 1.369e-06 ***
Residuals 45 101.004    2.245
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> coefcorel1 <- cov(arm,mort)/sqrt(var(arm)*var(mort))
> coefcorel1
[1] 0.6385633
```





Pour le lien entre la population et le nombre de morts par armes à feu, on obtient les résultats suivants.

call:
lm(formula = pop ~ mort)

Residuals:

	Min	1Q	Median	3Q	Max
	-84702404	-79684072	-74200629	-38749656	1232366102

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	86691636	54908832	1.579	0.121
mort	-881448	19669540	-0.045	0.964

Ce tableau permet donc de dire que $\beta_0 = 86691636$; $\beta_1 = -881448$

Le test de $H_0 : \beta_0 = 0$ contre $H_1 : \beta_0 \neq 0$. On a $t_0 = 86691636/54908832 = 1.579$.
valeur-p= 0,121, donc on rejette H_0 .

Le test de $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$. On a $t_0 = -881448/19669540 = -0.045$.
Puisque valeur-p= 0.964, on garde H_0 .

Residual standard error: 256900000 on 45 degrees of freedom
Multiple R-squared: 4.462e-05, Adjusted R-squared: -0.02218
F-statistic: 0.002008 on 1 and 45 DF, p-value: 0.9645

```
> confint(Reglin2)
          2.5 %      97.5 %
```

```
(Intercept) -23900429 197283701
mort -40497935 38735039
```

```
> anova(Reglin2)
```

Analysis of Variance Table

Response: pop

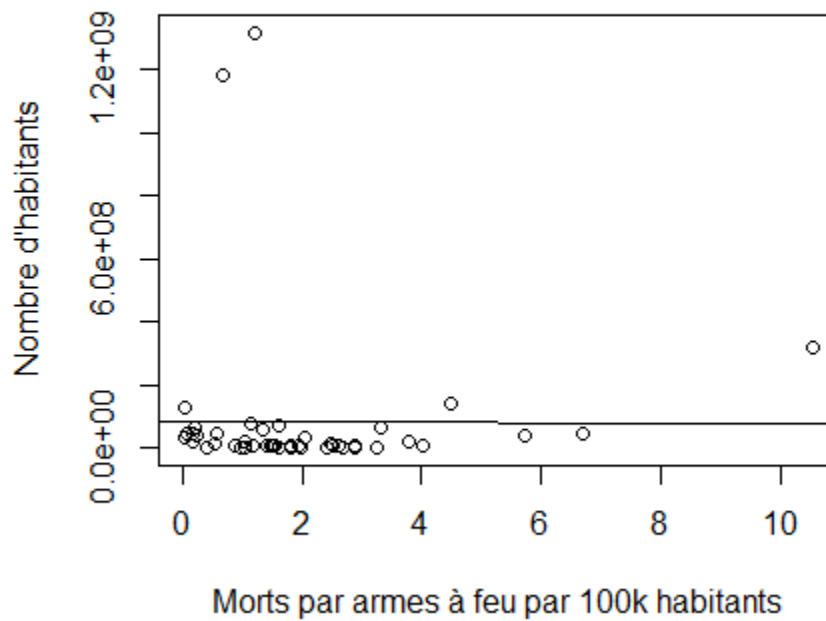
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mort	1	1.3251e+14	1.3251e+14	0.002	0.9645
Residuals	45	2.9692e+18	6.5983e+16		

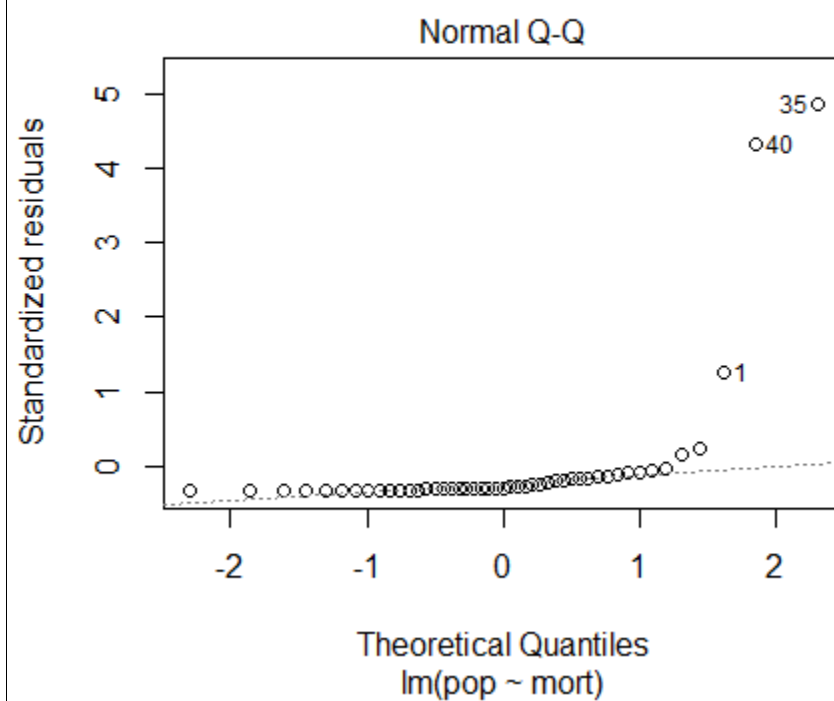
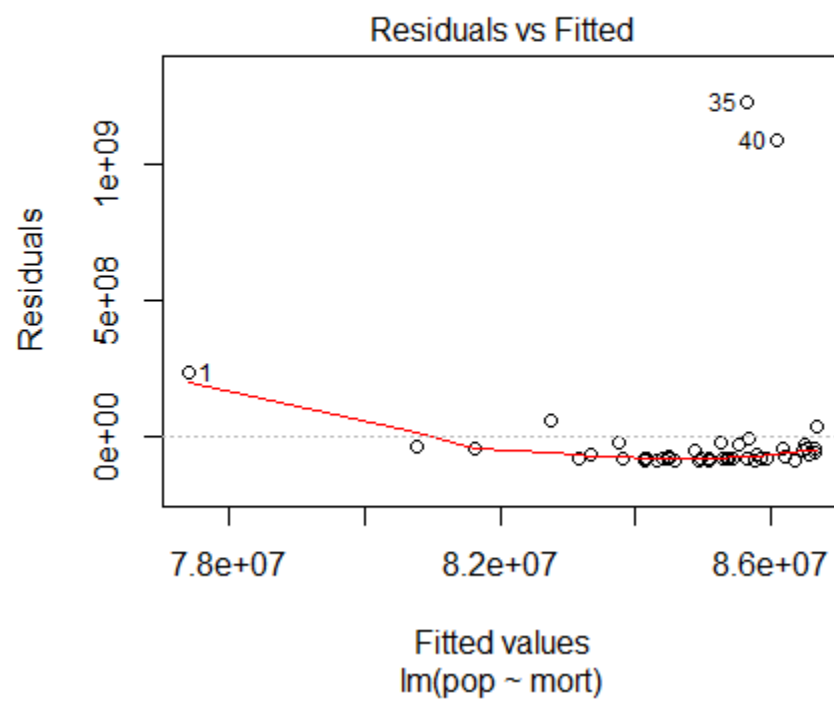
```
> coefcorel2 <- cov(mort,pop)/sqrt(var(mort)*var(pop))
```

```
> coefcorel2
```

```
[1] -0.006680154
```

Morts selon la population du pays





Conclusion

On a donc effectué des régressions linéaires avec les deux paires de variables qui étaient ciblées. Pour l'étude de la corrélation entre l'année de prise des données et le nombre de morts, il nous manque de données suffisantes pour pouvoir tirer des conclusions claires. En effet, en observant l'histogramme, on remarque que la majorité des données proviennent de la même année et qu'il n'y a qu'un petit nombre d'information les autres années. Il ne serait donc pas possible de faire d'analyse éclairée sur les différences entre les années.

Ensuite, par rapport au lien entre le nombre d'armes à feu et de morts, on suggère donc le modèle suivant à partir d'une régression linéaire : $Morts = 0.90132 + 0.07272 * \text{armes}$. De plus, on obtient un coefficient de corrélation moyen à fort de 0.64, ce qui nous démontre donc que 64% de la variation de la quantité de morts aux armes à feu est due à la variation de la quantité d'armes à feu dans le pays grâce à une relation linéaire positive entre les variables. En observant les graphiques des résidus, on voit que les points sont répartis aléatoirement autour de la ligne ce qui nous permet de confirmer qu'un modèle linéaire est approprié. On peut donc conclure qu'il existe en effet une corrélation entre l'accès aux armes à feu dans un pays et le nombre de victimes d'armes à feu.

Pour le lien possible entre le nombre de victimes d'armes à feu et la population d'un pays, on suggère aussi d'essayer un modèle par régression linéaire. Avec R, on obtient simplement une droite puisque la valeur-p1 est trop élevée pour obtenir un β_1 non-nul. On obtient donc un coefficient de corrélation quasi-nul de -0.007. De plus, comme on le voit sur les graphiques des résidus, leurs valeurs sont extrêmement élevées. On peut donc déduire qu'il n'existe aucune relation linéaire entre le nombre de victimes et le nombre d'habitants d'un pays.

En conclusion, nos données ont permis de prouver qu'il existe une corrélation entre l'accessibilité aux armes à feu et le nombre de victimes d'armes à feu par capita. Elles ont aussi permis de montrer qu'il n'existe pas de lien entre le nombre d'habitants d'un pays et le nombre de victimes. Cependant, les données étaient trop incomplètes pour permettre de prouver quoi que ce soit par rapport au temps et il faudrait donc des données de sources temporelles plus variées afin d'étudier plus en profondeur ce sujet. On ne peut que souhaiter que dans le futur, la quantité d'armes à feu disponibles va être réduite, et donc par corrélation, le nombre de victimes d'armes à feu sera aussi réduit.