FACULTÉ DES SCIENCES
UNIVERSITÉ D'ANGERS

Université d'Angers

Master Mathématiques et Applications

Master 2 DATA SCIENCE

Année académique 2020-2021

Projet d'Économétrie de l'évaluation

Analyse et prévision des tirages du jeu Euromillion

Étudiants:
Philippine RENAUDIN
Amadou Lindor FALL
Bertrand Carlos ADODO

Responsable du cours:
Philippe COMPAIRE

20 janvier 2021

Table des matières

Analyse et prévision des tirages du jeu Euromillions

0.1 Introduction

Ce rapport est le rendu du projet du cours **d'Économétrie de l'évaluation**. Les méthodes d'évaluations sont aujourd'hui très fréquemment utilisées pour estimer l'effet des interventions dans différents domaines. Les techniques disponibles sont certes multiples, mais il est important de rappeler que leurs spécificités et leurs hypothèses conditionnent fortement les résultats.
L'objectif de ce projet est de modéliser un problème économétrique et d'en analyser et interpréter les résultats en se basant sur une base de données, afin de répondre à un besoin de prévision.
Dans cette étude, nous sommes amenés à travailler sur une base de données contenant diverses informations sur les tirages du jeu **Euromillion**, notamment les boules et étoiles gagnantes. Nous avons choisi de mettre en application les méthodes vues au *chapitre 6* concernant **les modèles de comptage** en nous focalisant sur les 2 étoiles. C'est dans cette perspective que nous avons défini notre problématique d'étude : Quelles sont les probabilités que :

— **les deux étoiles soient supérieures ou égales à 10**

— **une seule soit supérieure ou égale à 10**

— **aucune des 2 étoiles soient supérieures ou égales à 10**

Pour cela, il va nous falloir construire notre variable endogène Y, qui prendra les modalités $0, 1$ ou 2. Puis, il nous faudra choisir les variables explicatives pour notre modèle. Enfin, après avoir estimés les paramètres du modèle et calculé les probabilités, nous conclurons sur les différentes valeurs obtenues et essaierons de déterminer les variables les plus influentes sur la valeur des étoiles.

0.2 Théorie du modèle de comptage

Notons que dans un modèle de comptage, la variable endogène Y est une variable qualitative ayant plusieurs modalités. La particularité de ce modèle est que les modalités, codées par des chiffres, représentent le nombre d'occurrences d'un évènement. Par exemple, on peut s'intéresser en économétrie au nombre d'enfants d'un couple, un nombre d'élèves par classe, un nombre de voitures par foyer etc. . .. Les variables exogènes quant à elles peuvent être quantitatives comme qualitatives.

Ceci étant fixé, nous détaillons dans la suite les mathématiques sous-jacentes au modèle. Notons que la meilleure loi de probabilité pour modéliser un nombre d'occurrence est **la loi de Poisson**. Donc dans les modèles de comptage, nous considérerons toujours que la variable endogène Y suit une loi de Poisson.

On rappelle la formule de la densité de la loi de Poisson :

Soit $Y \sim \mathbb{P}(\lambda)$, alors, $\mathbb{P}(Y = k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$, où $k \in \mathbb{N}$ et $\lambda > 0$ représente le nombre moyen d'occurrence d'un évènement.

Pour estimer $p_k = \mathbb{P}(Y = k)$, il faut donc réussir à estimer λ. Pour cela, on utilise la fonction génératrice des moments de la loi de Poisson et on obtient :

$$\mathbb{E}(Y) = \lambda = \exp(\beta_0 + \beta X), \text{ où :}$$

— X est le tableau de données constitué des individus en ligne et des variables exogènes en colonne (de taille $n \times p$)
— β_0 est la constante du modèle,

— $\beta = (\beta_,, \beta_2, \ldots, \beta_p)$ est le vecteur des coefficients de la régression

Or pour estimer λ, on voit qu'il faut connaitre β_i pour tout $i \geqslant 0$. Mais ces coefficients ne sont pas connus par définition. On va donc les estimer par la méthode du maximum de vraisemblance.

Une fois $\widehat{\beta_0}$ et $\widehat{\beta}$ déterminés, on peut donc calculer $\widehat{\lambda}$ et $\widehat{p_k}$ par les formules suivantes :

$$\widehat{\lambda} = \exp(\widehat{\beta_0} + \widehat{\beta} X) \qquad \widehat{p_k} = \frac{\widehat{\lambda}^k e^{-\widehat{\lambda}}}{k!}$$

Remarque : Dans ce modèle, les coefficients $\widehat{\beta_i}$ peuvent s'interpréter (pour les variables quantitatives uniquement).

0.3 Explication et transformation de la base de données

Avant de passer à l'application des modèles de comptage , il a fallu transformer la base de données (construction de la variable endogène Y) et sélectionner les variables explicatives (exogènes) à garder.
Notre base de données **Euromillion** recense les résultats des tirages du jeu Euromillion en Europe ainsi que diverses informations sur ces tirages. Elle est constituée de 175 observations et 53 variables. [https://github.com/Phi-gif/Projet_eco/blob/main/data/BDD_EUROMILLION.xlsx]

0.3.1 Définition de la variable endogène Y

Le principe est le suivant :

— Si aucune des 2 étoiles (E1, E2) n'est une dizaine, alors Y prend la modalité 0 : $Y = 0$
— Si exactement une des 2 étoiles (E1, E2) est une dizaine, alors Y prend la modalité 1 : $Y = 1$
— Si les 2 étoiles (E1, E2) sont des dizaines, alors Y prend la modalité 2 : $Y = 2$

0.3.2 Encodage des variables exogènes

Jour de tirage

Les tirages de l'Euromillion se font les mardi et vendredi. Nous devons donc transformer la variable **Jour de tirage** en variable binaire :

— 0 pour Mardi
— 1 pour Vendredi

Date de tirage

La variable **Date de tirage** du jeu de données couvre des dates allant de 2019 à 2020. On décide de garder uniquement l'année. Cette variable prendra ainsi 2 modalités que l'on code de la façon suivante :

— 0 pour l'année 2019
— 1 pour l'année 2020

Numéro de tirage dans le cycle

Cette variable indique la 'position' du tirage considéré par rapport au dernier tirage gagnant. Un cycle commence dès que le tirage précédent a été gagnant. Dans la base de données, les modalités 1 à 5 représentent presque la moitié de nos observations. On décide alors, pour faire diminuer le nombre de modalités de cette variable, de séparer les observations en 2 classes et on recode la variable de la façon suivante :

— 0 si le numéro de tirage est compris entre 1 et 5 (inclus)
— 1 sinon (incluant le Super Jackpot)

Boules

Nous avons également voulu utiliser l'information apportée par les boules. Dans le jeu de données, les boules, nommées B1-B5 sont numérotés de 1 à 50. On les recode selon le principe suivant :

— Si la somme des 5 boules est paire, la variable exogène 'boule' prendra la modalité 0
— Si la somme des 5 boules est impaire, la variable exogène 'boule' prendra la modalité 1

Choix des variables exogènes

Une fois les variables précédentes recodées, nous décidons d'ajouter la variable **nombre de grilles jouées** aux variables exogènes. Nous nous contenterons de garder ces 5 variables explicatives dans la suite. En effet, nous avions au préalable effectué une Analyse en Composantes Principales afin de déterminer les variables les plus significatives. Malheureusement, rien de probant n'en est ressorti, nous avons donc décidé de garder les variables qui nous paraissaient les plus pertinentes.

0.4 Interprétations des coefficients et des probabilités

0.4.1 Approche globale

Nous décidons dans un premier temps de travailler avec la base de données dans son ensemble (tous les tirages et les variables exogènes sélectionnées), avant de regarder séparément ce qu'il se passe pour l'année 2019 dans un premier temps, puis pour l'année 2020 dans un second temps. En effet, l'année 2020 étant assez spéciale, nous tenterons de voir si le Coronavirus a eu un quelconque effet sur l'Euromillion et donc sur les estimateurs et probabilités que nous allons calculer.

Nous décidons de transformer la variable 'nombre de grilles jouées' en la passant au logarithme dans un soucis de lisibilité des résultats et de facilité d'interprétation. Pour rappel, les variables constituant la base de données d'étude sont : le jour de tirage, l'année du tirage, la parité de la somme des boules de B1 à B5, le nombre de grilles jouées, le numéro de tirage dans le cycle et la variable endogène Y. Tout cela est résumé dans le graphique suivant :

Liste alphabétique des variables et des attributs

#	Variable	Type	Long.	Format	Libellé
4	Y	Num.	8	BEST.	Y
3	annee	Num.	8	BEST.	annee
1	boule	Num.	8	BEST.	boule
5	jour_de_tirage	Num.	8	BEST.	jour_de_tirage
6	nombre_de_grilles_jou_es	Num.	8	BEST.	nombre de grilles jouées
2	numero_de_tirage_dans_le_cycle	Num.	8	BEST.	numero_de_tirage_dans_le_cycle

Figure 1 – Variables retenues

Statistiques de la variable quantitative

La seule variable quantitative dans notre étude est le nombre de grilles jouées. Nous obtenons ici quelques informations statistiques sur cette variable (après passage au log) qui nous seront utiles pour la suite. Nous remarquons une valeur moyenne pour cette variable de 16,9465.

Statistiques sur les variables quantitatives (passées au log)

Procédure UNIVARIATE
Variable : ln_nb_grilles (ln_nb_grilles)

Moments			
N	175	Somme des poids	175
Moyenne	16.9464807	Somme des observations	2965.63411
Ecart-type	0.31955281	Variance	0.102114
Skewness	0.49642755	Kurtosis	0.82809413
Somme des carrés non corrigée	50274.829	Somme des carrés corrigée	17.767836
Coeff Variation	1.88565885	Std Error Mean	0.02415592

Mesures statistiques de base			
Emplacement		Variabilité	
Moyenne	16.94648	Ecart-type	0.31955
Médiane	16.94053	Variance	0.10211
Mode	16.98204	Intervalle	1.81864
		Ecart interquartile	0.40123

Remarque : Le mode affiché est le plus petit des 2 modes avec un effectif de 2.

Tests de tendance centrale : Mu0=0				
Test	Statistique		P-value	
t de Student	t	701.5456	Pr > |t|	<.0001
Signe	M	87.5	Pr >= |M|	<.0001
Rang signé	S	7700	Pr >= |S|	<.0001

Figure 2 – Mesures statistiques

Nous pouvons donc maintenant afficher le début de la base de données. La variable **année** n'est volontairement pas affichée ici car l'année n'a pas d'importance dans cette section.

Informations sur la nouvelle base de données créée (avec pasage au ln)

Obs.	boule	numero_de_tirage_dans_le_cycle	Y	jour_de_tirage	ln_nb_grilles
1	1	0	2	0	17.2580
2	1	0	1	1	16.8272
3	0	0	0	0	17.1641
4	1	1	0	1	16.9040
5	0	1	1	0	17.2336
6	1	0	1	1	16.6691
7	0	0	1	0	17.0207
8	0	0	0	1	16.7295
9	1	0	1	0	17.0803
10	1	0	0	1	16.8196

Figure 3 – Premières observations de la base de données

Estimation des paramètres

Avec notre base de données modifiée, nous pouvons enfin estimer les paramètres du modèle pour pouvoir ensuite calculer les probabilités. Nous utilisons le logiciel SAS pour effectuer cela. Pour pouvoir estimer les $\widehat{\beta}_i$ du modèle et les différents $\widehat{\lambda}$, nous lançons le code suivant.

```
/* PARTIE PREDICTION */
/* Ajout de lignes supplémentaires pour la prédiction de lambda */
data more;
input Y ln_nb_grilles boule jour_de_tirage numero_de_tirage_dans_le_cycle;
keep Y ln_nb_grilles boule jour_de_tirage numero_de_tirage_dans_le_cycle;

datalines;
. 16.9465 1 1 1
. 16.9465 1 1 0
. 16.9465 1 0 0
. 16.9465 0 0 0
. 16.9465 0 0 1
. 16.9465 0 1 1
. 16.9465 0 1 0
. 16.9465 1 0 1

;/* 16.9465 données trouvées par UNIVARIATE (les dummy c'est à nous de choisir) */

data newprojet2;
set projet2 more;
run;
```

Figure 4 – Code permettant d'ajouter des lignes dans la base de données à des fins de prédiction

Nous voyons donc que, pour les lignes supplémentaires, la valeur du Y est laissée vide car celle-ci va être prédite. La valeur du nombre de grilles jouées est fixée à la valeur moyenne pour tous les nouveaux individus. En revanche, les valeurs des 3 variables qualitatives sont à tour de rôle, soit égale à 0 ou 1, de façon à ce que toutes les situations possibles soient décrites. Nous allons donc obtenir 8 valeurs différentes pour $\widehat{\lambda}$ et 24 probabilités, correspondant à $\widehat{p_0}$, $\widehat{p_1}$, $\widehat{p_2}$ dans chaque situation.
Regardons d'abord la sortie suivante :

Résultats estimés des paramètres

Paramètre	DDL	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|
Intercept	1	4.847627	8.251946	0.59	0.5569
ln_nb_grilles	1	-0.343567	0.487449	-0.70	0.4809
boule	1	0.131804	0.236926	0.56	0.5780
jour_de_tirage	1	0.164548	0.271013	0.61	0.5437
numero_de_tirage_dans_le_cycle	1	-0.073296	0.291718	-0.25	0.8016

Figure 5 – Estimation des paramètres du modèle

Au vu de cette sortie, on remarque qu'aucune des variables explicatives n'est significative. Cela pourrait dire que notre choix de variable n'est pas bon et qu'il faudrait considérer d'autres variables. Cependant, si l'on réussissait à trouver des variables très significatives dans cette étude, cela voudrait dire qu'il est possible de trouver un modèle fiable prédisant le

nombre d'étoiles supérieures ou égale à 10 et donc qu'il est possible de trouver une 'combine' pour gagner plus souvent et plus facilement à l'Euromillion. Or, si on avait connaissance de ce qui influe sur l'Euromillion, nous serions déjà tous millionnaires. Il est donc tout à fait normal ici que les variables ne soient pas significatives.

Beaucoup de paramètres sont liés à des variables qualitatives et ne sont donc pas interprétables. Nous pouvons uniquement interpréter le paramètre lié à la variable quantitative, à savoir le nombre de grilles jouées et l'intercept. Nous pouvons affirmer de ce fait que l'augmentation de 1 % du nombre de grilles jouées entraine une diminution de 0.343 du nombre de boules supérieures ou égales à 10, TCEPA.
De plus, la constante du modèle vaut 4.85, ce qui veut dire que si toutes les autres variables sont nulles, alors il y aura plus de 4 étoiles supérieures ou égales à 10. Bien entendu, cela n'a pas de sens ici car il n'y a que 2 étoiles à l'Euromillion. Cette valeur n'est que mathématique et correspond à la constante du modèle.

Une fois ces paramètres estimés, il est alors possible d'estimer, à leur tour, les différents λ pour chaque situation. Nous obtenons les résultats suivants :

Prévisions modèle de Poisson

Obs.	lambda_chap
176	0.47162
177	0.50749
178	0.43049
179	0.37733
180	0.35066
181	0.41338
182	0.44482
183	0.40006

Figure 6 – Estimation des lambda

Calcul des probabilités

Avec ces estimations, il est alors possible de calculer les différentes probabilités dans chaque situation. Nous résumons tout cela dans le tableau suivant.

Lambda estimé	a	Configuration	Probabilité
0.47162	0	111	0.624
0.47162	1	111	0.294
0.47162	2	111	0.069
0.50749	0	110	0.602
0.50749	1	110	0.306
0.50749	2	110	0.078
0.43049	0	100	0.650
0.43049	1	100	0.280
0.43049	2	100	0.060
0.37733	0	000	0.686
0.37733	1	000	0.259
0.37733	2	000	0.049
0.35066	0	001	0.704
0.35066	1	001	0.247
0.35066	2	001	0.043
0.41338	0	011	0.661
0.41338	1	011	0.273
0.41338	2	011	0.057
0.44482	0	010	0.641
0.44482	1	010	0.285
0.44482	2	010	0.063
0.40006	0	101	0.670
0.40006	1	101	0.268
0.40006	2	101	0.054

Figure 7 – Tableau récapitulatif des probabilités de chaque situation

Une première constatation est que les λ estimés sont tous très proches de 0. Donc en moyenne, ce sont entre 0.3 et 0.5 étoiles qui sont supérieures ou égales à 10. On comprend bien que dans ce cas, cela revient à dire qu'en moyenne, 0 étoile n'est supérieure ou égale à 10. Une explication pourrait venir de l'inégale répartition des modalités de Y dans ce jeu de données.

Par conséquent, pour chaque situation, on a $\widehat{p_0} > \widehat{p_1} > \widehat{p_2}$.

On rappelle que :
- La configuration est la suivante : **boule**, **jour de tirage**, **numero de tirage dans le cycle**
- Le codage des boules est le suivant : 0 si la somme des boules (B1-B5) est paire, 1 sinon
- Le jour de tirage est codé en 0 s'il s'agit d'un Mardi et en 1 s'il s'agit d'un Vendredi.
- Quant au numéro de tirage dans le cycle, il est codé en 0 si le numéro de tirage est compris entre 1 et 5, en 1 sinon

Ce tableau s'interprète comme suit : La probabilité que Y soit égale à 0 sachant la configuration 1 1 1 vaut 0.623, la probabilité que le Y soit égale à 1 vaut 0.294 et la probabilité que le Y soit égale à 2 est très petite, égale à 0.069.

On remarque que c'est la configuration 0 0 1, c'est à dire pour un tirage effectué le mardi, ayant pour numéro de tirage dans le cycle un nombre supérieur à 5, dont la somme des boules B1-B5 est paire, qui donne une probabilité de n'avoir 0 étoile supérieure ou égale à 10 la plus élevée, pour une participation moyenne à l'Euromillion. Par conséquent, c'est également cette configuration qui donne la probabilité que les 2 étoiles soient supérieures ou égales à 10 la plus faible.

On trouve des résultats opposés pour la configuration 1 1 0, c'est à dire pour un tirage effectué le vendredi, ayant pour numéro de tirage dans le cycle un nombre inférieur ou égal à 5, dont la somme des boules B1-B5 est impaire, pour une participation moyenne.

Nous ne considérons que la valeur moyenne du nombre de grilles jouées dans la suite donc nous ne précisons plus 'pour une participation moyenne à l'Euromillion' à partir de maintenant.

0.4.2 Année 2019

Nous nous intéressons maintenant uniquement aux tirages ayant eu lieu en 2019. En effet, l'année 2020 étant assez particulière, il nous parait judicieux d'étudier séparément les résultats de l'Euromillion pour ces deux années. Nous filtrons donc les lignes de notre base de données pour ne garder que les tirages de 2019 (sur SAS). Nous reprenons la même démarche que précédemment.

Statistiques de la variable quantitative

D'après le graphique suivant, la valeur moyenne du nombre de grilles jouées, après passage au logarithme est maintenant de 17.05. Il y avait donc plus de grilles jouées en moyenne en 2019 que sur les 2 années combinées. Nous pourrons comparer cette valeur à celle de 2020 pour identifier un potentiel 'effet coronavirus'.

Mesures statistiques de base			
Emplacement		Variabilité	
Moyenne	17.04699	Ecart-type	0.30314
Médiane	16.99168	Variance	0.09190
Mode	16.98204	Intervalle	1.31740
		Ecart interquartile	0.35546

Figure 8 – Mesures statistiques

Estimation des paramètres

Nous lançons cette fois-ci le code modifié suivant :

```
/* PARTIE PREDICTION */
/* Ajout de lignes supplémentaires pour la prédiction de lambda */
data more;
input Y ln_nb_grilles boule jour_de_tirage numero_de_tirage_dans_le_cycle;
keep Y ln_nb_grilles boule jour_de_tirage numero_de_tirage_dans_le_cycle;

datalines;
. 17.0470 1 1 1
. 17.0470 1 1 0
. 17.0470 1 0 0
. 17.0470 0 0 0
. 17.0470 0 0 1
. 17.0470 0 1 1
. 17.0470 0 1 0
. 17.0470 1 0 1
```

Figure 9 – Code permettant d'ajouter des lignes dans la base de données à des fins de prédiction

Nous obtenons les estimations β_i suivantes :

Estimation des paramètres bêta_i (interprétables)

The COUNTREG Procedure

Model Fit Summary	
Dependent Variable	Y
Number of Observations	88
Missing Values	8
Data Set	WORK.NEWPROJET2019
Model	Poisson
Log Likelihood	-71.38922
Maximum Absolute Gradient	2.27968E-6
Number of Iterations	4
Optimization Method	Newton-Raphson
AIC	152.77844
SBC	165.16512

Algorithm converged.

Résultats estimés des paramètres					
Paramètre	DDL	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|
Intercept	1	2.486136	13.539559	0.18	0.8543
ln_nb_grilles	1	-0.191065	0.795526	-0.24	0.8102
boule	1	0.289947	0.345314	0.84	0.4011
jour_de_tirage	1	-0.201523	0.396004	-0.51	0.6108
numero_de_tirage_dans_le_cycle	1	-0.406811	0.429955	-0.95	0.3441

Figure 10 – Estimation des paramètres du modèle

Les paramètres liés aux variables qualitatives ne sont toujours pas interprétables. De plus, nos variables sont toujours non significatives. Ouf !
Ici, la constante du modèle vaut 2.48 donc comme auparavant, si toutes les autres variables étaient nulles, on aurait plus de 2 étoiles supérieures ou égales à 10. Encore une fois, cette valeur n'a pas vraiment de sens ici.
L'augmentation de 1 % du nombre de grilles jouées a pour effet de diminuer de 0.19 le nombre d'étoiles supérieures ou égales à 10, ce qui est plus faible que précédemment.

On remarque que le critère AIC est passé de 294 (étude sur les 2 années), à 152 (étude sur 2019). On peut donc penser que notre modèle « fit » mieux l ?année 2019 que les 2 années combinées.

Nous pouvons donc maintenant estimer les λ comme précédemment. Nous regroupons les estimations et les probabilités dans un même tableau récapitulatif.

Calcul des probabilités

Lambda estimé	a	Configuration	Probabilité
0.33644	0	111	0.714
0.33644	1	111	0.240
0.33644	2	111	0.040
0.50534	0	110	0.603
0.50534	1	110	0.305
0.50534	2	110	0.077
0.61816	0	100	0.539
0.61816	1	100	0.333
0.61816	2	100	0.103
0.46257	0	000	0.630
0.46257	1	000	0.291
0.46257	2	000	0.067
0.30797	0	001	0.735
0.30797	1	001	0.226
0.30797	2	001	0.035
0.25176	0	011	0.777
0.25176	1	011	0.196
0.25176	2	011	0.025
0.37814	0	010	0.685
0.37814	1	010	0.259
0.37814	2	010	0.049
0.41155	0	101	0.663
0.41155	1	101	0.273
0.41155	2	101	0.056

Figure 11 – Tableau récapitulatif des probabilités de chaque situation

Les λ estimés sont toujours très proches de 0 même si cette fois, l'un d'eux dépasse 0.6. Donc on a toujours qu'en moyenne, 0 étoile est supérieure ou égale à 10.

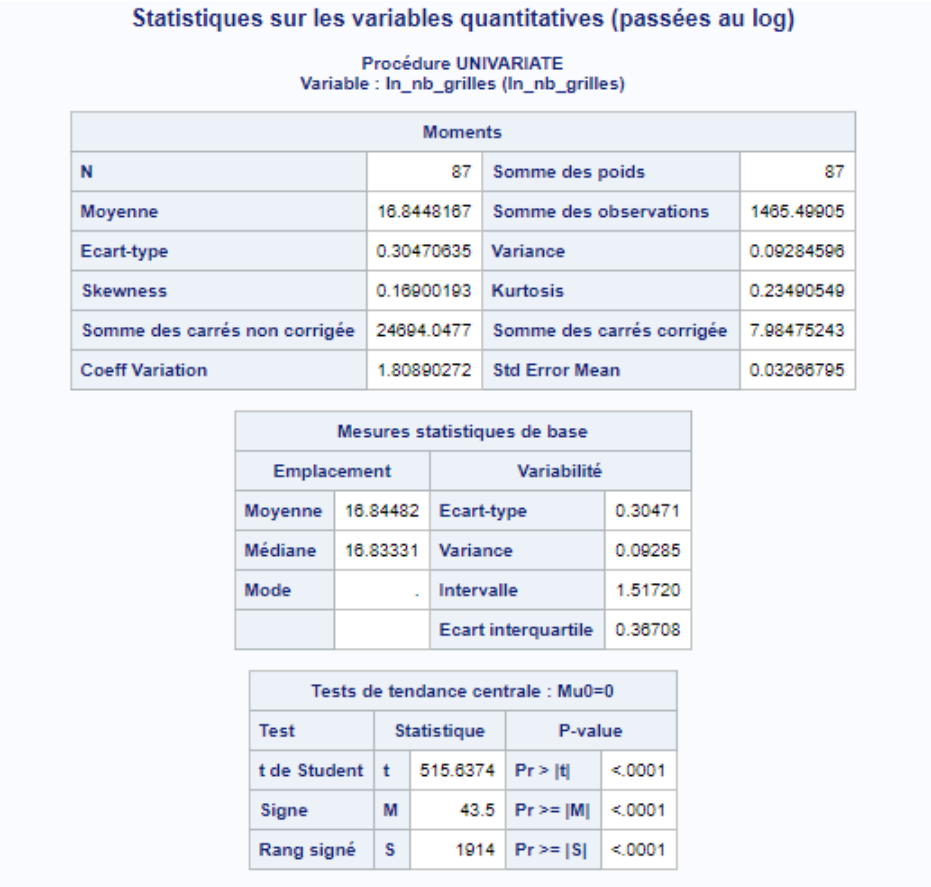
Par conséquent, pour chaque situation, on a toujours $\widehat{p_0} > \widehat{p_1} > \widehat{p_2}$.

Les configurations remarquables sont cette fois-ci la configuration 0 1 1 et la configuration 1 0 0. Pour l'année 2019, le tirage qui donnait la plus grande probabilité qu'aucune des deux étoiles ne soit supérieure ou égale à 10 était : un tirage le vendredi, dont le numéro de tirage dans le cycle est supérieur à 5 et dont la somme des boules B1-B5 est paire. Le tirage qui donnait la plus grande probabilité que les deux étoiles soient supérieures ou égales à 10 était : un tirage le mardi, dont le numéro de tirage dans le cycle est inférieur ou égal à 5 et dont la somme des boules B1-B5 est impaire.

0.4.3 Année 2020

Statistiques de la variable quantitative

D'après le graphique suivant, la valeur moyenne du nombre de grilles jouées, après passage au logarithme est maintenant de 16.84. Il y a donc eu beaucoup moins de grilles jouées en moyenne en 2020 qu'en 2019. Cela peut tout à fait s'expliquer par la pandémie qui nous a tous forcés à rester le plus possible chez nous. La vente de grille d'Euromillion en a donc très certainement été affectée négativement. De plus, dans un contexte de crise comme celui-ci, le pouvoir d'achat diminue et donc il est aussi assez naturel que moins de personnes dépense de l'argent dans des jeux d'argent.



Statistiques sur les variables quantitatives (passées au log)

Procédure UNIVARIATE
Variable : ln_nb_grilles (ln_nb_grilles)

Moments			
N	87	Somme des poids	87
Moyenne	16.8448167	Somme des observations	1465.49905
Ecart-type	0.30470635	Variance	0.09284596
Skewness	0.16900193	Kurtosis	0.23490549
Somme des carrés non corrigée	24694.0477	Somme des carrés corrigée	7.98475243
Coeff Variation	1.80890272	Std Error Mean	0.03266795

Mesures statistiques de base			
Emplacement		Variabilité	
Moyenne	16.84482	Ecart-type	0.30471
Médiane	16.83331	Variance	0.09285
Mode	.	Intervalle	1.51720
		Ecart interquartile	0.36708

Tests de tendance centrale : Mu0=0				
Test	Statistique		P-value	
t de Student	t	515.6374	Pr > |t|	<.0001
Signe	M	43.5	Pr >= |M|	<.0001
Rang signé	S	1914	Pr >= |S|	<.0001

Figure 12 – Mesures statistiques

Estimation des paramètres

Nous lançons cette fois-ci le code modifié suivant :

```
/* PARTIE PREDICTION */
/* Ajout de lignes supplémentaires pour la prédiction de lambda */
data more2;
input Y ln_nb_grilles boule jour_de_tirage numero_de_tirage_dans_le_cycle;
keep Y ln_nb_grilles boule jour_de_tirage numero_de_tirage_dans_le_cycle;

datalines;
. 16.8448 1 1 1
. 16.8448 1 1 0
. 16.8448 1 0 0
. 16.8448 0 0 0
. 16.8448 0 0 1
. 16.8448 0 1 1
. 16.8448 0 1 0
. 16.8448 1 0 1
```

Figure 13 – Code permettant d'ajouter des lignes dans la base de données à des fins de prédiction

Nous obtenons les estimations β_i suivantes :

Estimation des paramètres bêta_i (interprétables)

The COUNTREG Procedure

Model Fit Summary	
Dependent Variable	Y
Number of Observations	87
Missing Values	8
Data Set	WORK.NEWPROJET2020
Model	Poisson
Log Likelihood	-68.48088
Maximum Absolute Gradient	1.56636E-7
Number of Iterations	4
Optimization Method	Newton-Raphson
AIC	146.96176
SBC	159.29130

Algorithm converged.

Résultats estimés des paramètres

Paramètre	DDL	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|
Intercept	1	4.820551	11.303582	0.43	0.6698
ln_nb_grilles	1	-0.360921	0.669951	-0.54	0.5901
boule	1	0.024810	0.340416	0.07	0.9419
jour_de_tirage	1	0.581089	0.394377	1.47	0.1406
numero_de_tirage_dans_le_cycle	1	0.298478	0.395106	0.76	0.4500

Figure 14 – Estimation des paramètres du modèle

Les variables sont toujours non significatives. Ici, la constante du modèle vaut 4.82. Elle est très proche de la constante du premier modèle. De plus, l'augmentation de 1 % du nombre de grilles jouées a pour effet de diminuer de 0.36 le nombre d'étoiles supérieures ou égales à 10, ce qui est plus élevé que précédemment et proche des résultats du premier modèle.

Le critère AIC est passé de 152 (2019), à 146 pour l'année 2020. Notre modèle a donc l'air de mieux « fitter » l'année 2020 que dans les deux cas précédents.

Nous pouvons donc maintenant estimer les λ comme précédemment. Nous regroupons les estimations et les probabilités dans un même tableau récapitulatif.

Lambda estimé	a	Configuration	Probabilité
0.70137	0	111	0.496
0.70137	1	111	0.348
0.70137	2	111	0.122
0.52038	0	110	0.594
0.52038	1	110	0.309
0.52038	2	110	0.080
0.29104	0	100	0.747
0.29104	1	100	0.218
0.29104	2	100	0.032
0.28391	0	000	0.753
0.28391	1	000	0.214
0.28391	2	000	0.030
0.38266	0	001	0.682
0.38266	1	001	0.261
0.38266	2	001	0.050
0.68419	0	011	0.504
0.68419	1	011	0.345
0.68419	2	011	0.118
0.50763	0	010	0.602
0.50763	1	010	0.306
0.50763	2	010	0.078
0.39227	0	101	0.676
0.39227	1	101	0.265
0.39227	2	101	0.052

Figure 15 – Tableau récapitulatif des probabilités de chaque situation

On remarque ici des valeurs de λ estimés très hétérogènes et ayant une amplitude bien plus grande que précédemment. On a toujours néanmoins $\widehat{p_0} > \widehat{p_1} > \widehat{p_2}$.

Les configurations remarquables sont cette fois-ci la configuration 0 0 0 et la configuration 1 1 1. Pour l'année 2020, le tirage qui donnait la plus grande probabilité qu'aucune des deux étoiles ne soit supérieure ou égale à 10 était : un tirage le mardi, dont le numéro de tirage dans le cycle est inférieur ou égal à 5 et dont la somme des boules B1-B5 est paire. Le tirage qui donnait la plus grande probabilité que les deux étoiles soient supérieures ou égales à 10 était : un tirage le vendredi, dont le numéro de tirage dans le cycle est supérieur à 5 et dont la somme des boules B1-B5 est impaire.

0.5 Conclusion

Dans cette étude, nous avons donc pu créer plusieurs modèles afin de répondre à la problématique posée. Comme évoqué dans les sections précédentes, les variables explicatives considérées ne sont pas significatives. Pour autant, dans cette situation, cela est tout à fait normal. Les probabilités des différentes modalités de la variable endogène ont pu donner certaines informations quant aux situations favorisant un nombre d'étoiles supérieures ou égales à 10 élevé (ou faible selon ce qui nous intéresse le plus). Pour toutes les situations considérées, c'est la modalité 0 qui a la plus forte probabilité. Cela semble normal puisque la variable cible Y prend beaucoup plus souvent cette modalité dans le jeu de données. Une explication naturelle à ce phénomène est que dans le cadre de L'Euromillion, il y a beaucoup plus de chance d'obtenir une valeur d'étoile inférieure à 10 : en effet, la première étoile à 9 chances sur 12 d'être inférieure à 10, de même pour la deuxième. Donc au final, si la répartition des valeurs des étoiles est uniforme, on a à chaque tirage un peu plus de 56% de chance que la variable Y prenne la modalité 0, un peu plus de 37% de chance que la variable Y prenne la modalité 1 et un peu plus de 6% de chance que la variable Y prenne la modalité 2.
Les configurations intéressantes sont donc celles donnant des probabilités très différentes de celles mentionnées ci-dessus. Dans chaque approche considérée ici, nous avons pu remarquer des probabilités bien différentes de celles-ci. On peut donc se demander si, au final, certaines variables n'auraient pas une influence sur la valeur des étoiles.
Au vu de la répartition des tirages dans le jeu de données en fonction de l'année (autant de tirage datés de 2019 que de 2020), la différence de moyenne observée pour la variable **nombre de grilles jouées** pour l'année 2019 et l'année 2020 nous semble être essentiellement dû à la pandémie mondiale.
Enfin, nous avons considéré pour cette même variable sa valeur moyenne tout au long de l'étude mais il peut aussi être intéressant de regarder ce qu'il se passe pour les jours de tirage où la participation est grande, ou à l'inverse faible.