Codes et interprétations

Comme énoncé précédemment, nous travaillerons d'abord sur la base de données dans son ensemble, avant de regarder ce qui 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 répondre à la question suivante : as t il eu un effet corona virus

Tout d'abord, la présente base de données est modifiée avec des variables codés 0 1 et des passages au log dans le but d'avoir des ordres de grandeur plus réalistes

Les variables que nous retenons dans notre base de données sont les jours de tirages, les années de tirages, la parité de la somme des boules de b1 a b5, le nombre de grilles jouées, le numéro de tirage dans le cycle (codés en 0 ou 1) et notre variable endogène Y. Cette variable Y que nous tentons d'expliquer par rapport au variables précédentes fait référence aux Etoiles. Elle est codée de la façon suivante : si les 2 étoiles sont toutes les 2 inferieur a 9, si elles sont supérieures à 10 ou si l'une est inferieurs a 9 a l'autres supérieurs a 10

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

Ainsi, nous arrivons a tirer quelques informations de la variable quantitative (avant le passage au log) de notre base de données

Statistiques sur les variables quantitatives

Procédure UNIVARIATE
Variable : nombre_de_grilles_jou_es (nombre de grilles jouées)

Moments			
N	175	Somme des poids	175
Moyenne	24159875.5	Somme des observations	4227978221
Ecart-type	8685067.97	Variance	7.54304E13
Skewness	1.7340324	Kurtosis	4.02542222
Somme des carrés non corrigée	1.15272E17	Somme des carrés corrigée	1.31249E16
Coeff Variation	35.9483142	Std Error Mean	656529.427

Mesures statistiques de base			
Emplacement		Variabilité	
Moyenne	24159876	Ecart-type	8685068
Médiane	22760402	Variance	7.54304E13
Mode	23724920	Intervalle	51454101
		Ecart interquartile	9004224

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	36.79938	Pr > |t|	<.0001
Signe	M	87.5	Pr >= |M|	<.0001
Rang signé	S	7700	Pr >= |S|	<.0001

Ayant qu'une seule variable quantitative à savoir le nombre de grilles jouées, certaines de ses données sont connues à savoir sa moyenne, son écart type, sa variance etc…

Il est tout à fait normal que ces statistiques changent après le passage au log. Nous nous retrouvons avec des valeurs moins importantes en taille et plus facile à manier comme l'atteste l'image ci-dessous

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

Les nouvelles statistiques après le passage au log sont les suivantes :

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

Disposant de notre base de données modifiée, nous pouvons enfin commencer notre prédiction a l'aide de la recherche du lambda dans nos différents cas de figure.

```
/* 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;
```

Notre variable Y est codée comme étant le nombre de boules supérieur à 10. En effet, il en va de soi que la probabilité que Y soit égale à 0 fait référence à la probabilité que 0 boule soit supérieur à 10 . Les deux boules sont donc toutes les deux comprises entre les entiers 0 et 10. La probabilité que notre Y soit égale a 1 nous donne la probabilité qu'une boule soit supérieur à 10, la probabilité que Y soit égale à 2 fait référence a la probabilité que les 2 boules soient supérieures à 10. La problématique est restée la même mais elle a juste été reformulée différemment.

```
/* Estimation : Modèle de comptage (Poisson) */

proc countreg data=newprojet2; /* estime les beta et donne le niveau de significaivité des variables associées */
model Y = ln_nb_grilles boule jour_de_tirage numero_de_tirage_dans_le_cycle / dist=poisson;
output out=outestim pred=lambda_chap ; /*pred = lambda chapeau, pas les probas */

title 'Estimation des paramètres bêta_i (interprétables)';
run;
```

Estimation des paramètres bêta_i (interprétables)

The COUNTREG Procedure

Model Fit Summary	
Dependent Variable	Y
Number of Observations	175
Missing Values	8
Data Set	WORK.NEWPROJET2
Model	Poisson
Log Likelihood	-142.26526
Maximum Absolute Gradient	1.12811E-6
Number of Iterations	4
Optimization Method	Newton-Raphson
AIC	294.53052
SBC	310.35445

Algorithm converged.

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

Interprétation : *Les paramètres ne sont pas du tout significatifs. On s'en doutait un peu car si on avait connaissance de ce qui influe sur l'euro million, le nombre de gagnants aurait fortement augmenté.

Certains paramètres sont positifs, d'autres sont négatifs. Leurs signes témoignent de leurs apports positifs ou négatifs sur l'explication de notre variable Y. Cependant, du fait de leurs caractères qualitatifs, ces betas ne sont pas interprétables, nous essaierons simplement d'interpréter les betas de la variable quantitatif à, savoir le nombre de grilles. Nous pouvons affirmer de ce fait que l'augmentation de 1% de ce nombre entraine une diminution de 0.343 du nombre de boules supérieurs à 10.

De plus, pour chaque configuration et pour un nombre de grille joué en moyenne égale à 16.9465, nous avons un lambda chapeau, lambda chapeau qui nous permettra de calculer la probabilité sachant que les étoiles soient en dessous de 10, une au-dessus et les deux au-dessus (a=0 , 1 et 2). Le lambda chapeau apporte de plus une certaine information par rapport à sa proximité aux entiers 0, 1 et 2. Un lambda estimé très proche de 0 nous montre que, en moyenne, les boules sont très proches de 0. Sachant que c'est souvent le cas, nous comprenons ceci comme étant le fait que, très souvent, 0 boule est supérieur à 10 . Ceci est justifiée par une illégale répartition de notre jeu de données avec les quasis ¾ des observations codées comme 0 en ce qui concerne la somme des boules.

Résumons ceci dans le tableau ci-dessous

Lambda estimé	a	Proba	Congiguration	
0.47162	0	0.6239905842933695	1 1 1	
0.47162	1	0.29428643936443893	1 1 1	
0.47162	2	0.06939568526652834	1 1 1	
0.50749	0	0.6020047158898076	1 1 0	
0.50749	1	0.30551137326691846	1 1 0	
0.50749	2	0.07752198340961423	1 1 0	
0.43049	0	0.6501904233477652	1 0 0	
0.43049	1	0.2799004753469794	1 0 0	
0.43049	2	0.060247177816060586	1 0 0	
0.37733	0	0.6856897589356568	0 0 0	
0.37733	1	0.25873131673919136	0 0 0	
0.37733	2	0.04881354387259954	0 0 0	
0.35066	0	0.7042231490268047	0 0 1	****
0.35066	1	0.24694288943773937	0 0 1	
0.35066	2	0.04329649680511884	0 0 1	
0.41338	0	0.6614108989263401	0 1 1	
0.41338	1	0.2734140373981705	0 1 1	
0.41338	2	0.056511947389827864	0 1 1	
0.44482	0	0.6409396347839807	0 1 0	
0.44482	1	0.2851027683446103	0 1 0	
0.44482	2	0.06340970670752478	0 1 0	
0.40006	0	0.6702798280394291	1 0 1	
0.40006	1	0.268152148005454	1 0 1	
0.40006	2	0.05363847416553097	1 0 1	

Tout d'abord , apportons une explication a la configuration . Cette configuration joue sur la logique suivante : `boule jour_de_tirage numero_de_tirage_dans_le_cycle ;`

Pour rappel, en ce qui concerne le codage des boules, ce codage prend 0 si la somme des boules (variables exogènes B1 à B5) est pair .Elle est codée 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 et en 1 sinon.

Ce tableau s'interprète comme suit : avec le premier lambda, la probabilité que Y soit égale à 0 sachant la configuration 1 1 1 est 0.623, la probabilité que le Y soit égale a 1 est de 0.069 et la probabilité que le Y soit égale a 2 est très petite, égale à 0.069. Ces prévisions sont valables pour la base de données dans son ensemble. Dans la suite, nous traiterons la prévision suivant l'année 2019 dans un premier temps, puis suivant l'année 2020 dans un second temps.

2019

Statistiques sur les variables quantitatives (passées au log)

Procédure UNIVARIATE
Variable : ln_nb_grilles (ln_nb_grilles)

Moments			
N	88	Somme des poids	88
Moyenne	17.0469893	Somme des observations	1500.13506
Ecart-type	0.30314281	Variance	0.09189557
Skewness	1.01807515	Kurtosis	0.81204442
Somme des carrés non corrigée	25580.7813	Somme des carrés corrigée	7.99491426
Coeff Variation	1.77827773	Std Error Mean	0.03231513

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

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	527.5234	Pr > |t|	<.0001
Signe	M	44	Pr >= |M|	<.0001
Rang signé	S	1958	Pr >= |S|	<.0001

Nous restons toujours avec la base au format log car comme dit précédemment, elle est plus facile à manier. Le tableau ci-dessus nous témoigne des statistiques de la variable quantitative (nombre de grilles jouées) . Notre étude utilisera encore une fois essentiellement, la moyenne, qui est de 17.0470 grilles jouées. Ainsi, à l'aide de cette donnée, nous pouvons commencer nos prédictions pour une configuration donnée.

```
/* 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
```

```
/* Estimation : Modèle de comptage (Poisson) */

proc countreg data=newprojet2019; /* estime les beta et donne le niveau de significaivité des variables associées */
model Y = ln_nb_grilles boule jour_de_tirage numero_de_tirage_dans_le_cycle / dist=poisson;
output out=outestim2019 pred=lambda_chap ; /*pred = lambda chapeau, pas les probas */

title 'Estimation des paramètres bêta_i (interprétables)';
run;
```

L'estimation des bêtas, comme fait précédemment , est donné dans le tableau ci-dessous. Les betas des variables qualitatives ne sont toujours pas interprétables. Nous interprèterons uniquement le beta de la variable quantitative après avoir discuté de la significativité des variables explicatives .

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

Les coefficients des variables explicatives restent non significatifs. Nous pouvons dire que notre modèle fait fi du fait que l'année soit précisée ou pas. La conclusion reste la même : « Si on connaissait quelles variables jouent sur l'euromillion , le nombre gagnant aurait énormément augmenté » . Nous pouvons dire que l'augmentation de 1% du nombre e grilles joué entraine une diminution de 0.19 du nombre de boules supérieurs à 10.

Lambda estimé	a	Configuration	Proba	
0.33644	0	111	0.7143087408148703	
0.33644	1	111	0.24032203275975497	
0.33644	2	111	0.04042697235084598	
0.50534	0	110	0.6033004184100645	
0.50534	1	110	0.30487183343934204	
0.50534	2	110	0.07703196615511855	
0.61816	0	100	0.5389351665509708	
0.61816	1	100	0.3331481625551482	
0.61816	2	100	0.10296943408254519	****
0.46257	0	000	0.6296633295358314	
0.46257	1	000	0.2912633663433895	
0.46257	2	000	0.06736484768473083	
0.30797	0	001	0.7349373657993242	
0.30797	1	001	0.2263386605452179	
0.30797	2	001	0.03485275864405537	
0.25176	0	011	0.7774312991925221	****
0.25176	1	011	0.19572610388470935	
0.25176	2	011	0.024638001957007213	
0.37814	0	010	0.6851345751107226	
0.37814	1	010	0.2590767882323686	
0.37814	2	010	0.04898364835109393	
0.41155	0	101	0.6626223890467389	
0.41155	1	101	0.272702442121854	
0.41155	2	101	0.0561153043027624 5	

2020

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

```
/* 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
```

```
/* Estimation : Modèle de comptage (Poisson) */

proc countreg data=newprojet2020; /* estime les beta et donne le niveau de significaivité des variables associées */
model Y = ln_nb_grilles boule jour_de_tirage numero_de_tirage_dans_le_cycle / dist=poisson;
output out=outestim2020 pred=lambda_chap ; /*pred = lambda chapeau, pas les probas */

title 'Estimation des paramètres bêta_i (interprétables)';
run;
```

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

Lambda estimé	a	Configuration	Proba
0.70137	0	111	0.4959054477329505
0.70137	1	111	0.3478132038764595
0.70137	2	111	0.1219728734014162
0.52038	0	110	0.5942946730809127
0.52038	1	110	0.3092590619778453
0.52038	2	110	0.08046611533601557
0.29104	0	100	0.7474857779889749
0.29104	1	100	0.21754826082591125
0.29104	2	100	0.21754826082591125
0.28391	0	000	0.7528343967530161
0.28391	1	000	0.2137372135821488
0.28391	2	000	0.030341066154053933
0.38266	0	001	0.6820447550850001
0.38266	1	001	0.2609912459808261
0.38266	2	001	0.049935455093511466
0.68419	0	011	0.5044987080722391
0.68419	1	011	0.34517297107594525
0.68419	2	011	0.11808194754022548
0.50763	0	010	0.6019204411289539
0.50763	1	010	0.3055528735302909
0.50763	2	010	0.07755390259509079
0.39227	0	101	0.6755216984770551
0.39227	1	101	0.2649868966615944
0.39227	2	101	0.05197320497672182

A rajouter :

*Dire peut-être pourquoi les proba pour a=0 sont les plus fortes (done)

*Rappeler les configurations pour aider le lecteur (done)

*Faire pour 2020 (done)

*et y introduire effet corona

*Rappeler qu'en classe, on joue sur le passé ,on faisait une Verif alors que là,on prévoit le turfu

*Dire qu'au lieu de jouer sur la valeur moyenne du nombre de grilles joué, on peut jouer sur le nombre de grille max ou min par exemple

*Parler de la minimisation du critère AIC (comparer les critères AIC et BIC de 2019 et 2020) et dire le quel fit mieux notre jeu de données

*Faire les mêmes screens pour 2020 et 2019 en y rajoutant les moyennes , nbres de grilles joués (done)