POLYTECHNIQUE MONTRÉAL

Travail de session partie 2

MTH2302D - Probabilités et statistiques

Trimestre : Automne 2019

Alice GONG 1961605
Agnès SAM YUE CHI 1954192
Kai Sen TRIEU 1963091

Groupe: 04

Présenté à :
Charafeddine TALAL

Département de mathématiques et de génie industriel
École Polytechnique de Montréal
Lundi le 2 décembre 2019

Partie 1

Contexte général des données

Dans le cadre du travail de session du cours de *Probabilités et Statistiques*, il semblait intéressant de se pencher sur le sujet du cinéma, une forme de divertissement qui influence de nos jours grandement la culture populaire, avec les productions hollywoodiennes plus grandioses les unes que les autres. Avec l'émergence de tous les sites dont la thématique tourne autour du 7e art et tous les détails fournis sur chaque sortie de films, les données sur ceux-ci sont abondantes et permettent de former des liens entre différents paramètres. Ceci peut être utile pour évaluer les caractéristiques d'un film qui pourraient entraîner sa popularité ou son déclin. La grande quantité d'informations fournies sur ce sujet nous a permis de choisir une base de données. Les sources dont elle provient et les variables qui la composent seront présentées dans les prochaines sections, ainsi que les liens que nous pouvons observer entre les variables obtenues.

Provenance des données

Les données proviennent originellement du site Internet Movie Database (IMDb), qui rassemble les informations de classement et de production sur des millions de films. Les bases de données peuvent être trouvées dans le lien suivant: https://www.imdb.com/interfaces/.
Toutefois, nous avons utilisé les données fournies par un utilisateur de Kaggle, un site de compétition de sciences de données, car les informations sont plus accessibles, visibles et organisées. De plus, on y trouve le score (ou Metascore) des films provenant du site de Metacritic, qui représente le score accordé par des critiques professionnels. Les données sont présentées sommairement sous le lien suivant: https://www.kaggle.com/isaactaylorofficial/imdb-10000-most-voted-feature-films-041118/version/1.

Description de la forme des données

Les données recueillies fournissent de l'information concernant les 10 000 films les mieux classés sur IMDb. Pour ne pas alourdir le fichier, nous avons limité les données à 3 000 films, ce qui respecte le minimum de 50 observations. Originellement, le tableau contenant l'information à propos de ces films est classé sous forme de onze colonnes/variables. Par contre, en raison de leur impertinence, certains variables ont été omises. Celles que nous avons retenues et qui seront utilisées dans le cadre de ce travail sont l'année de sortie, le score, le Metascore, le nombre de votes sur IMDb, le genre et le revenu de chaque film. Six variables seront donc traitées dans l'analyse.

Nous avons appris au début de la session qu'il existe plusieurs type de variables aléatoires, dont la variable aléatoire discrète, et la variable aléatoire continue. Une variable est dite discrète, lorsque sa valeur se retrouve dans l'ensemble des entiers relatifs. Au contraire, une variable ne peut qu'être considéré comme étant continue lorsqu'elle appartient à l'ensemble des nombres réels. Sachant cela, nous avons classé nos variables ainsi:

Variable aléatoire discrète

- **Metascore** : score, allant de 0 à 100, attribué à un film par plusieurs critiques professionnels de publications renommées. Ces données sont recueillies du site *metacritic.com*;
- **Année**: année de sortie du film;
- **Nombre de votes**: nombre d'utilisateurs de IMDb ayant évalué ce film en votant pour un score allant de 0 à 10. Le rang des films sur IMDb est en fonction de cette variable;
- **Genre**: catégorie narrative ou émotionnelle du film traité. Nous avons décidé d'associer chaque film à un chiffre, de sorte que nous pourrions identifier le genre que nous voulons étudier par ce nombre. De plus, il arrive que plus d'un genre est associé à un film. Nous avons décidé de garder le premier genre de chaque film puisque c'est certainement le plus pertinent et va faciliter notre étude.

Variable aléatoire continue

- **Score**: score de 0 à 10 alloué à un film selon un vote effectué par les utilisateurs de IMDb. C'est une évaluation donnée par le grand public;
- **Revenu**: calculé en million de dollar US.

Questions ouvertes

Les questions en gras ont été modifiées depuis la première remise dans le but d'obtenir une analyse plus diversifiée. Une des questions a aussi été retirée, car elle était similaire à une autre en termes d'analyse descriptive.

1. La relation entre le nombre de votes et le revenu est-elle linéaire?
2. Quel(s) genre(s) de films est le plus populaire au fil des années (par décennies)?
3. Peut-on établir des liens entre le genre de film et le revenu accumulé?
4. Est-ce que le score a une corrélation avec le Metascore?

Partie 2

Analyse des données

Chacune des analyses réalisées aura pour but de répondre aux questions ouvertes posées dans la partie 1 et vous pouvez utiliser différentes techniques pour répondre à la même question. L'emploi de techniques non vues en cours est autorisé et encouragé, mais vous devrez alors les expliquer et fournir des liens et des références.

Avant d'entamer l'analyse de nos données, il importe de mentionner que nous avons dû effectuer un nettoyage de ces derniers au préalable, car nous avons réalisé que certaines informations n'étaient pas nécessaires pour maintes raisons. Plus précisément,, nous avons d'abord retiré tous les films ayant des données vides ou n'ayant pas de valeur (NA), car celles-ci se révèlent aberrantes dans les analyses. Nous nous sommes permis ce type de filtration, puisque nous avons un nombre considérable de données. Cela a réduit le nombre d'observations de 237 données et le genre de film "Thriller" a été éradiqué de l'échantillon. Étant donné que l'échantillon est d'une taille assez considérable, cette suppression aura un effet minime sur nos tests.

Statistique descriptive

Vue globale: Test de Pearson

Nous avons utilisé le test de khi-deux de Pearson pour déterminer la corrélation entre chacune des deux variables, sous forme de *heat map*. Comme nous avons vu lors du cours, ce test permet de déterminer s'il existe une association/dépendance entre deux variables quelconques.

Si nous nous se basons sur la figure suivante, les variables ayant le plus grand potentiel de dépendance seraient le score et le Metascore. Puisque la corrélation entre le nombre de votes et le revenu est supérieur à 0.5, il est aussi possible de produire une régression linéaire avec ces deux variables.

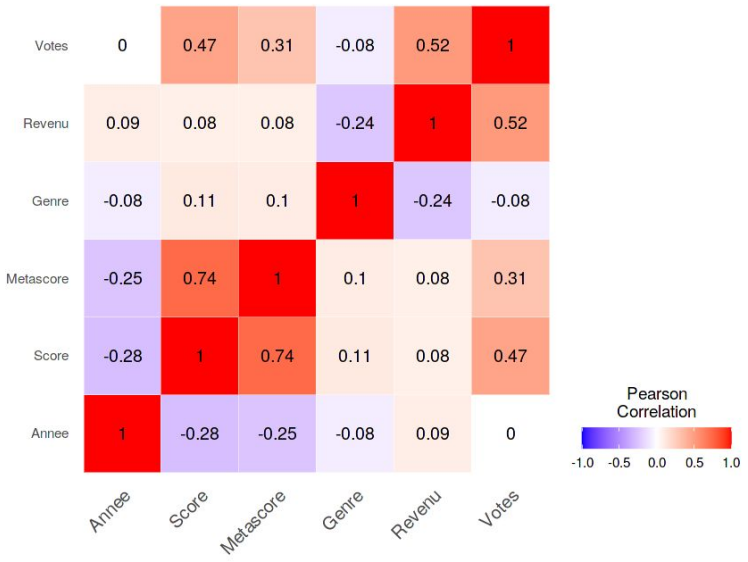


Figure 1. Heat map pour illustrer la corrélation Pearson entre les variables listées.

1. Relation entre le nombre de votes et le revenu

Un nuage de points a été choisi pour représenter la relation entre le nombre de votes et le revenu. Ce type de diagramme est idéal pour visualiser facilement une corrélation entre deux variables. Toutefois, une grande dispersion de valeurs peut mener à un graphique moins clair et donc rendre la tâche de déterminer la corrélation plus ardue. De plus, si une variable est de type catégorie comme le genre de film, un nuage de point n'est pas possible, car les valeurs seraient représentées sous formes de lignes verticales parallèles. Pour répondre à la question ci-présente, c'est donc un diagramme adapté.

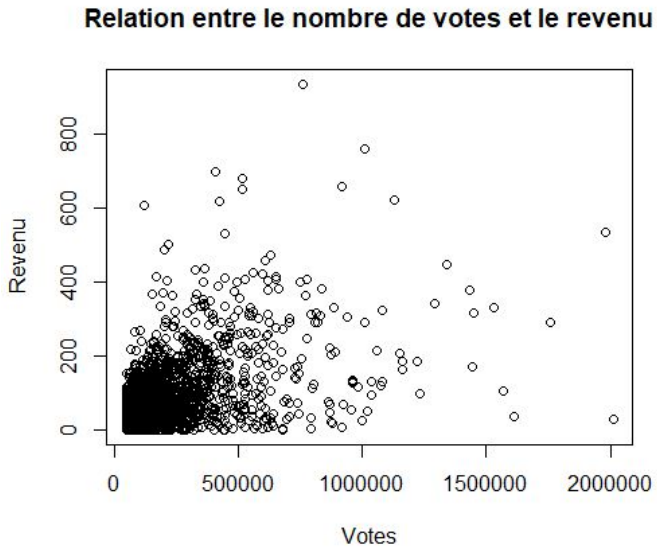


Figure 2. Nuage de points représentant la relation entre les variables du revenu et le nombre de votes.

2. Popularité d'un genre par rapport à l'année de sortie

Les questions impliquant le genre de film suivent le groupement suivant:

Liste de genres: Action (1), Adventure (2), Animation (3), Biography (4), Comedy (5), Crime (6), Drama (7), Family (8), Fantasy (9), Film noir (10), Horror (11), Musical (12), Mystery (13), Romance (14), Sci-fi (15), Western (16)

Groupe 1- Action, adventure, fantasy, western, sci-fi

Groupe 2- Horror, crime, film noir, drama, mystery

Groupe 3- Family, animation, comedy

Groupe 4- Biography, musical, romance

Pour déterminer quel genre était le plus est plus populaire au fil des années, l'utilisation d'histogrammes est le choix le plus judicieux. En effet, ce diagramme permet la séparation en intervalles et de déterminer la fréquence d'une variable désirée. Les intervalles en question sont la variable des années séparée en décennie et l'histogramme déterminera la fréquence de chaque groupe de genres. Les autres types de diagrammes, comme les diagrammes de boite à moustache ou les nuage de points, sera peu utile pour répondre à la question puisqu'ils ne peuvent pas nous fournir des données sur la fréquence de chaque groupe de genres.

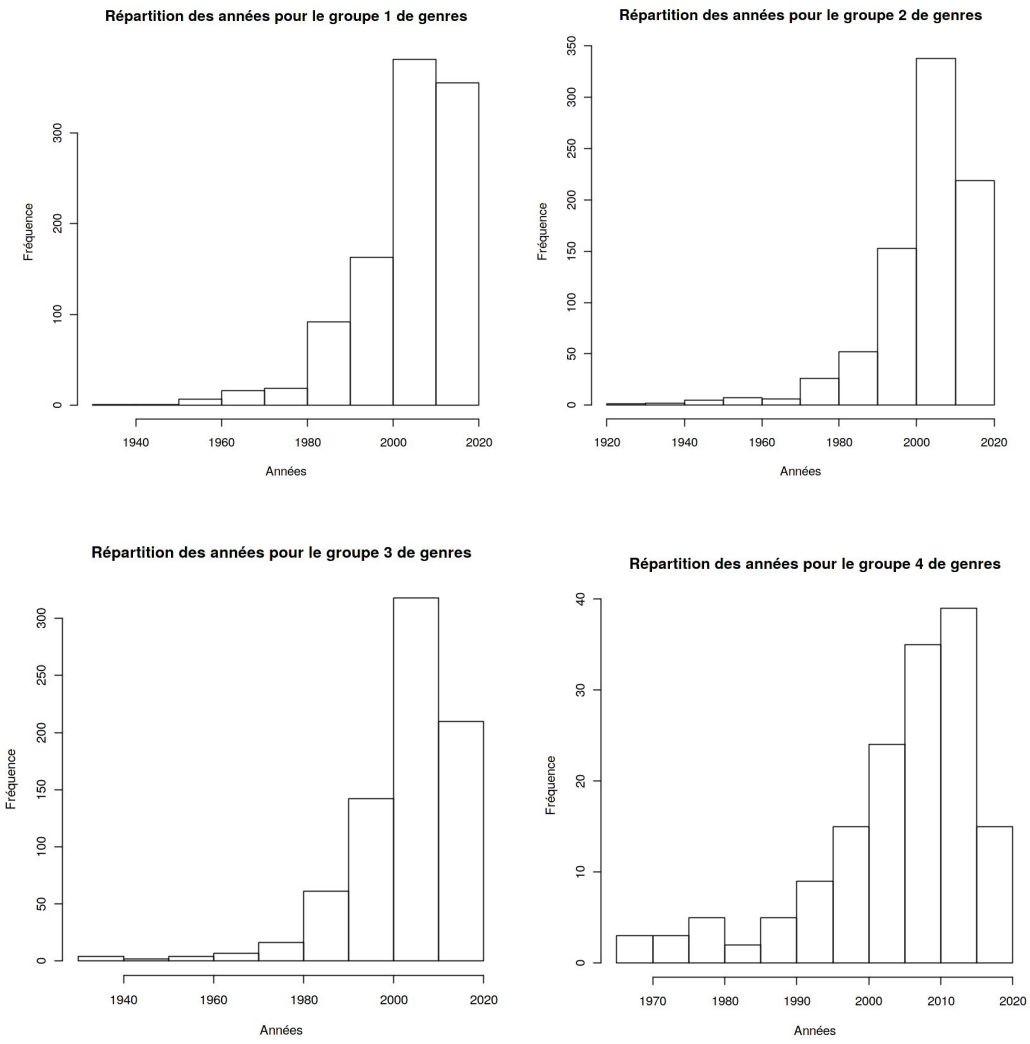


Figure 3. Répartition des genres de film de groupes différents à travers le temps.

Si nous comparons ces diagrammes entre eux, nous pouvons voir que la grande majorité des genres de films sont apparus après les années 2000. Il est possible de recenser plus de 300 films qui sont du genre des trois premiers groupes lors de la première décennie du 21e siècle. En ce qui concerne le groupe 4, leur popularité atteint les sommets qu'entre 2005 et 2015, mais seulement à un maximum de 40 films. Si nous observons l'allure générale des trois premiers graphique, nous remarquons qu'ils sont relativement similaires au fil des années. Cela signifie que le film le groupe de genre de films le plus populaires sont les films d'horreur, de crime, dramatique, mystères et les films noirs.

3. Relation entre le genre du film et le revenu perçu

Pour analyser le genre du film avec le revenu associé, une analyse préliminaire a été effectuée avec des boîtes à moustaches pour comparer ces relations entre les différents groupes de genres. Ce type de diagramme a été choisi pour mieux représenter chacun des groupes de genres, qui n'est originellement pas une valeur pouvant avoir des intervalles. Il aurait donc été impossible de visualiser ces données avec un nuage de points ou un histogramme, par exemple. Toutefois, les boîtes à moustache ne permettent pas de tirer des conclusions concrètes quant à la corrélation; il est seulement possible d'analyser la distribution de valeurs sélectionnées en fonction d'un paramètre d'une certaine valeur. Ici, par exemple, on analyse les données de revenus qui font partie de genres de groupe 1.

Un premier diagramme montre plusieurs données aberrantes se trouvant pour la plupart au dessus d'un revenu de 200 millions de dollars américains. Afin d'augmenter la visibilité, ces valeurs ont été retirées des boîtes à moustache.

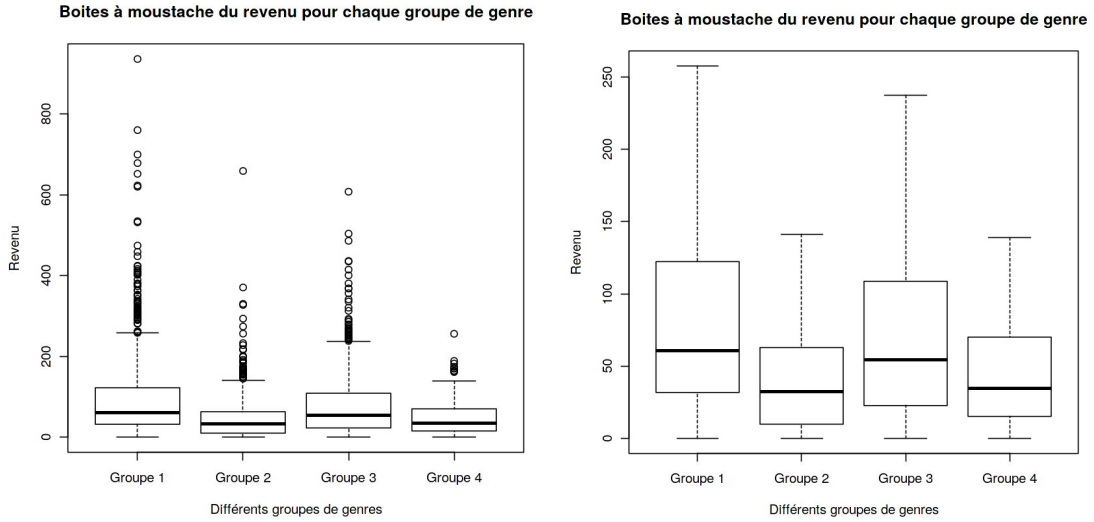


Figure 4. Corrélation entre le genre de films et le revenu.

Aucune des lignes médianes des diagrammes ne se trouvent au milieu de la boîte, signifiant qu'aucun des revenus des groupes ne suivent une distribution parfaitement symétrique. De plus, on observe des maximums et des boîtes beaucoup plus grandes pour les groupes 1 et 3 que pour les groupe 2 et 4, ce qui montre que la distribution des groupes 1 et 3 est plus dispersée et que 25% des films ayant un meilleur revenu pour ces groupes varient avec un plus grand intervalle.

4. Relation entre le score du film et le Metascore

Nous avons choisi d'effectuer une régression linéaire avec un nuage de points pour déterminer la relation entre le score, établi par des amateurs de films, et le Metascore, obtenu à partir de critiques professionnels. Ce type de graphique permet de visualiser rapidement le type de corrélation entre deux variables, en plus de montrer la présence de données aberrantes. Par contre, les nuages de points peuvent souvent avec une forme plutôt abstraite, ce qui rend difficile l'analyse de la corrélation. Dans ce cas-ci, la forme est assez clairement linéaire. Le coefficient de corrélation obtenu de 0.74 explique aussi que la linéarité explique bien le modèle. La régression est assez faiblement significative, puisque le modèle n'explique que 54,11% de la variance.

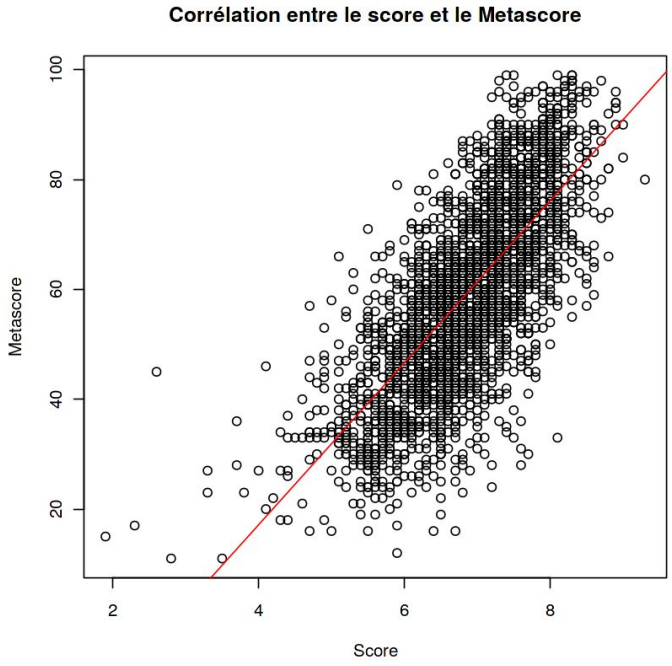


Figure 5. Nuage de points représentant la relation entre les variables du score et le metascore.

Estimation des paramètres

Pour déterminer la normalité, le test de Shapiro-Wilk sera utilisé dans les prochaines sections. Toutefois, pour une taille d'échantillon aussi grande que la nôtre, ce test, selon plusieurs statisticiens, aura toujours tendance à rejeter la normalité, malgré les évidences graphiques qui pointent directement à cette distribution [1]. On se basera donc sur des preuves supplémentaires telles que les Q-Q plots pour déterminer ou non une normalité.

1. Année

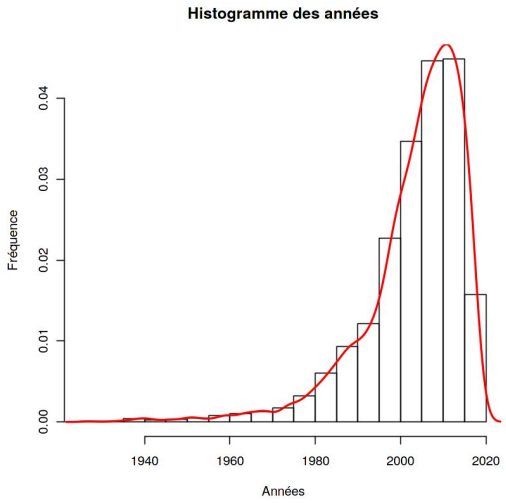


Figure 6. Diagramme illustrant le test de Shapiro-Wilk avec la variable « Années ».

Le test de Shapiro-Wilk donne une W de 0.84762 qui est plus proche de 0.7 que de 1, et une P-value de moins de 2.2e-16, qui est largement inférieur à une valeur de alpha de 0.05. L'année ne suit donc pas une distribution normale.

Un graphique de Cullen et Frey permet d'estimer le type de distribution le plus représentatif d'une série de données [2].

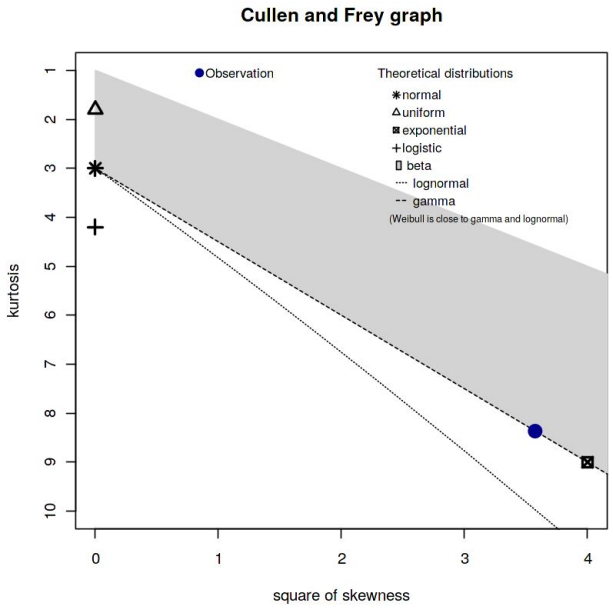


Figure 7. Graphique de Cullen et Frey pour la variable "Année".

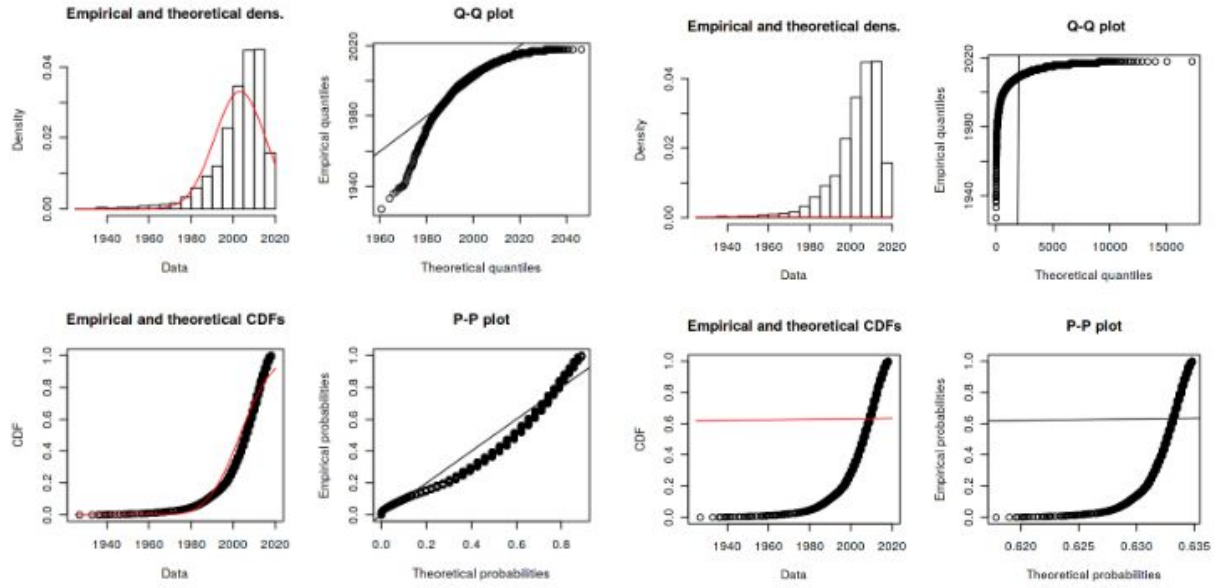


Figure 8. Tests d'ajustement pour une distribution gamma (les quatre diagrammes de gauche) et exponentielle (à droite).

Selon le graphique de Cullen et Frey, des distributions exponentielle et gamma seraient applicables pour l'analyse de la variable "Année" [2]. L'ajustement obtenu avec la distribution

gamma étant plus juste, celle-ci semble plus appropriée pour l'analyse. De plus, l'histogramme de la densité de la variable ressemble visuellement à une distribution gamma.

Nous pouvons estimer les paramètres de cette fonction gamma: la forme α vaut 2.778272e+04 et l'échelle λ, 7.207896e-02. L'intervalle de confiance se trouve entre les valeurs 2002.797 et 2003.698.

Pour le test d'hypothèse où l'hypothèse nulle est $\mu = \mu_0$, on utilise une moyenne calculée à partir de 1500 observations de moins que les données originales. On obtient un $t = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = 0.88285$ et une P-value de 0.3774. On ne rejette donc pas l'hypothèse nulle.

2. Score

En effectuant le test de normalité de Shapiro-Wilk, nous obtenons une statistique W de 0.98224, mais une P-value inférieure à 2.2e-16. Comme expliqué précédemment, nous allons tout de même considérer la normalité [1]. Avec un graphique de Cullen et Frey, nous pouvons voir que l'observation de nos données s'apparie à des distributions lognormales ou normales [2].

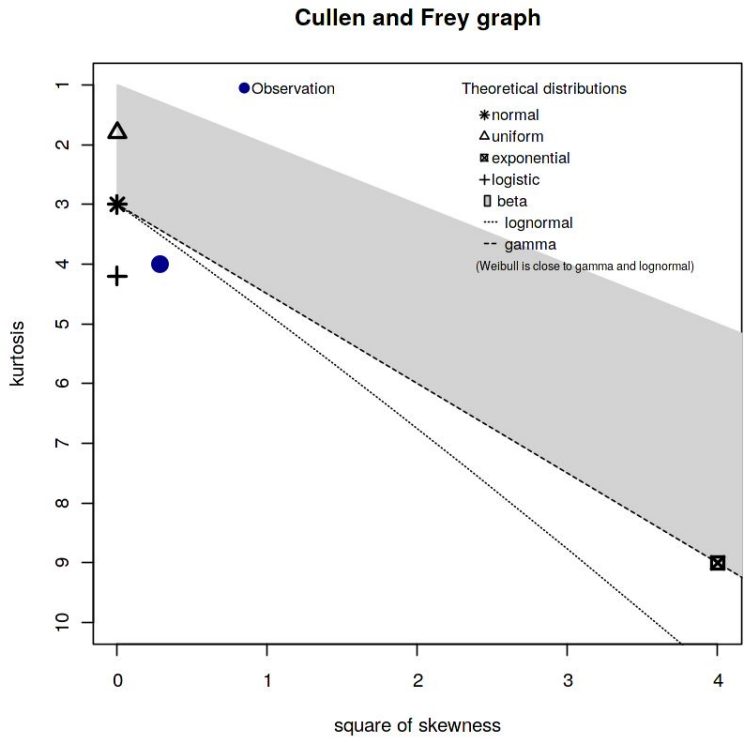


Figure 9. Graphique de Cullen et Frey pour la variable « Score ».

En testant l'ajustement de chaque distribution pour le score, on remarque que les deux sont acceptables, mais selon le Q-Q plot, la distribution normale semble plus adaptée. La normalité sera donc utilisée pour une analyse.

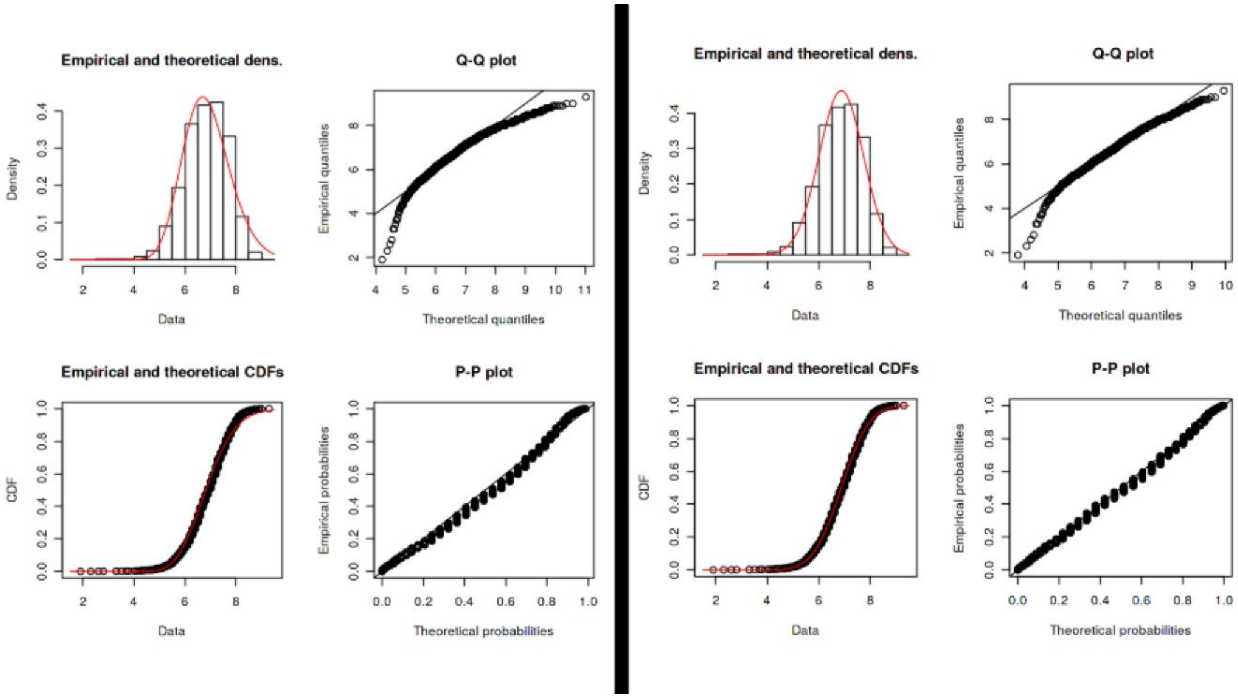


Figure 10. Tests d'ajustement pour une distribution lognormale (les quatre diagrammes de gauche) et normale (à droite).

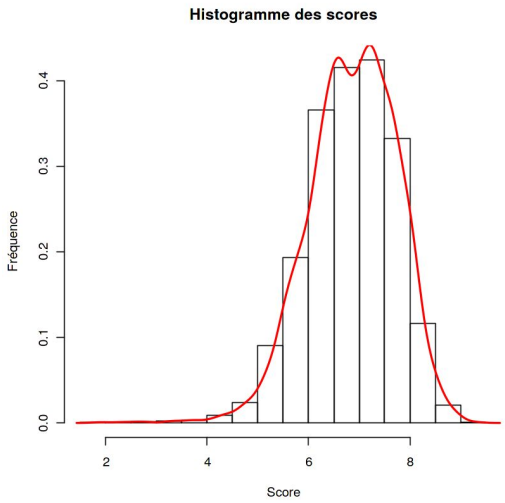


Figure 11. Diagramme illustrant le test de Shapiro-Wilk avec la variable « Score ».

On estime une moyenne de 6.87209555 et un intervalle de confiance entre les valeurs 6.22725281171497 et 7.51693828491912.

Le test d'hypothèse où l'hypothèse nulle est $\mu = \mu_0$ donne une valeur de t de 1.5424 et une P-valeur de 0.1231. On accepte l'hypothèse nulle.

3. Metascore

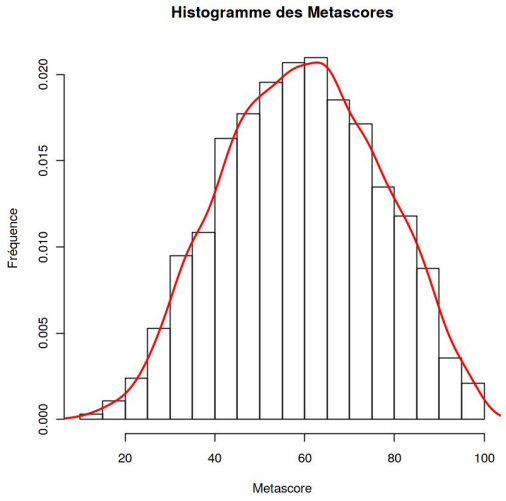


Figure 12. Diagramme illustrant le test de Shapiro-Wilk avec la variable « Metascore ».

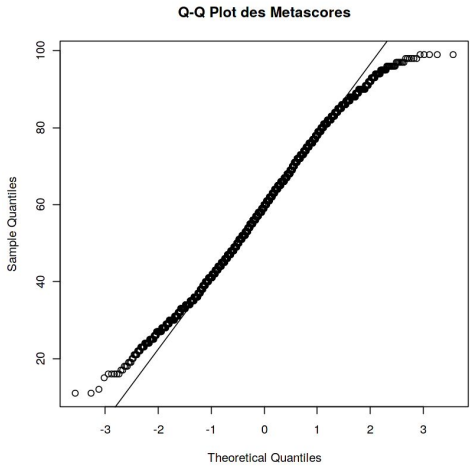


Figure 13. Diagramme de quartiles pour la variable « Metascore ».

L'histogramme de cette variable suggère visuellement une distribution normale. Pour confirmer l'étendue de la distribution normale sur cette variable, un diagramme de quartiles ainsi que la ligne représentant la normalité théorique a été produit. On remarque que la grande partie centrale suit parfaitement une distribution normale et que les données aux extrêmes dévient légèrement.

Si on évalue la normalité avec Shapiro-Wilk, on obtient un W de 0.93648, ce qui est en accord avec l'hypothèse de la normalité. La P-value obtenue est inférieure à 0.05 avec une valeur de 0.000345, mais comme expliqué précédemment, ce ne sera pas pris en compte pour notre choix de distribution [1].

Avec la méthode de vraisemblance, on estime ponctuellement la moyenne à 59.5085052. L'intervalle de confiance obtenu est [58.8636625113169, 60.153347984521].
Le test d'hypothèse où l'hypothèse nulle est $\mu = \mu_0$ donne une valeur de t de 0.55316 et une P-valeur de 0.5802. On accepte l'hypothèse nulle.

4. Genre

Afin de déterminer la distribution la plus représentatif pour la variable du genre, il serait judicieux d'utiliser le graphique de Cullen et Frey [2].

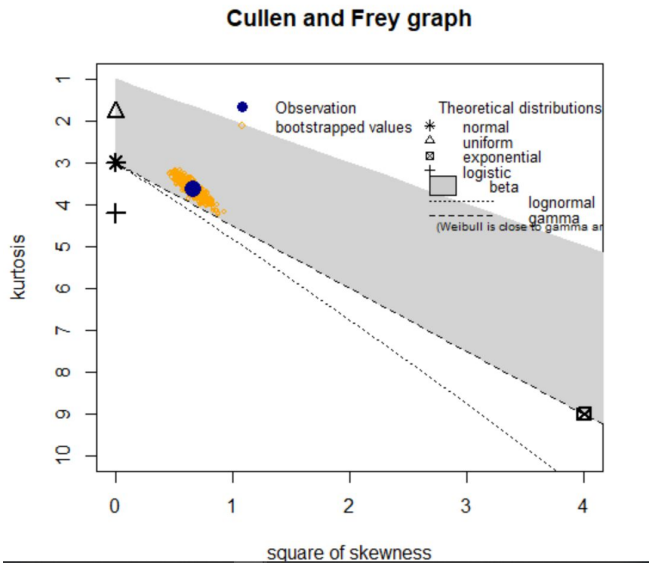


Figure 14. Graphique de Cullen et Frey pour la variable "Genre".

On voit à partir de ce graphique le genre suit une distribution beta. En connaissant cette information, il est possible de faire l'estimation ponctuelle de ses paramètres à l'aide de la méthode des moments. Nous trouvons donc alpha = -11.02333 et beta = 8.438901 et l'intervalle de confiance est entre 4.154693 et 4.375889.

En ce qui concerne le test d'hypothèse, la valeur obtenue pour p-value est 0.8339. Cette valeur est supérieur à 0.05, signifiant qu'il n'Est pas possible de rejeter l'hypothèse nulle.

5. Revenu

Suite au test de normalité de Shapiro-Wilk, nous avons obtenus les résultats suivants:

W = 0.73296, p-value < 2.2e-16

Nous pouvons donc rejeter l'hypothèse que le revenu suit une distribution normale puisque p-value est beaucoup plus petit que 0.05 et que alpha est plus proche de 0.7 que 1.

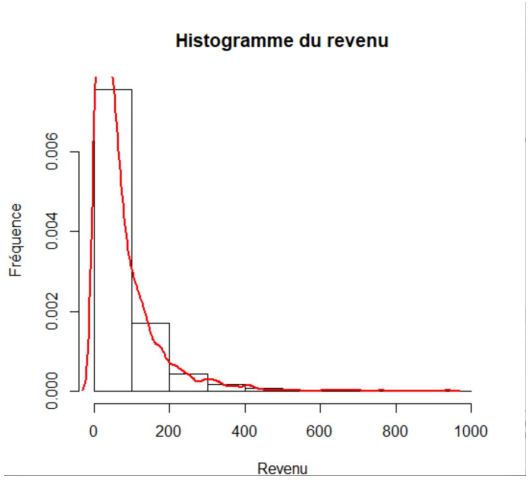


Figure 15. Diagramme illustrant le test de Shapiro-Wilk avec la variable « Revenu ».

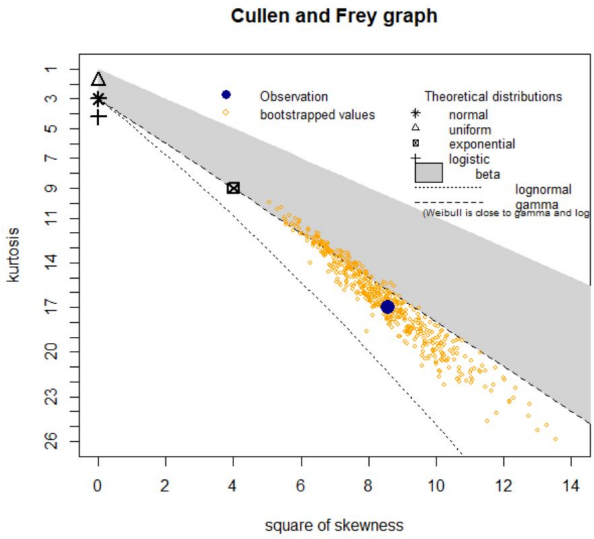


Figure 16. Graphique de Cullen et Frey pour la variable « Revenu ».

À l'aide du graphique de Cullen et Frey, il est possible de voir que la variable revenu suit une distribution gamma puisqu'elle tend vers cette dernière.
Avec la méthode du maximum de vraisemblance, nous obtenons une estimation de variable k = 0.7580868 et theta = 96.7442056.

Nous calculons par la suite l'intervalle de confiance qui se situe entre 70.19774 et 76.48326.
Pour ce qu'il en est du test d'hypothèse, nous trouvons une valeur p de 0.841 qui est beaucoup plus grand de 0.05. Nous ne pouvons donc pas rejeter l'hypothèse nulle.

6. Nombre de votes

Afin de déterminer si la variable du nombre de vote suit une distribution normale, nous avons utilisé le test de Shapiro-Wilk. Ce test nous informe que cette distribution donne W = 0.64244 et p-value < 2.2e-16. Étant donné que p-value de cette distribution est inférieur à 0.05 et que alpha est plus proche de 0.7 que 1, nous pouvons conclure qu'elle ne suit pas une distribution normale.

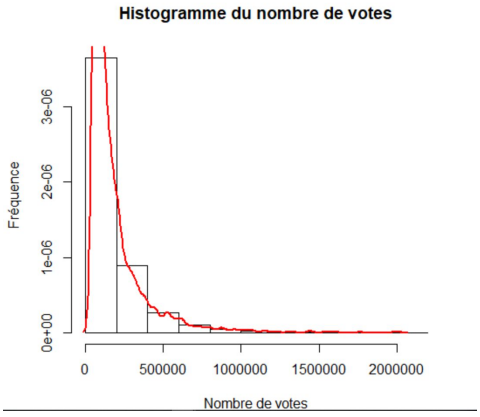


Figure 17. Diagramme illustrant le test de Shapiro-Wilk avec la variable « nombre de votes ».

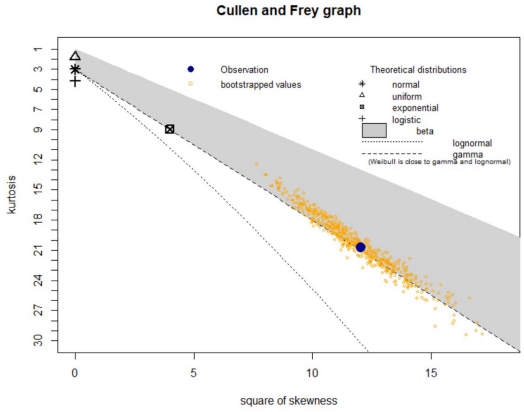


Figure 18. Graphique Cullen and Frey pour la variable « nombre de votes ».

Il est possible de déduire en générant le graphique Cullen and Frey que la variable nombre de votes suit une distribution gamma.

Il faut alors trouver la valeur des deux paramètres:

K = 1.722438 et theta = 1.051120e+05

En calculant l'intervalle de confiance, nous trouvons que cette intervalle se situe entre 173988.8 et 188109.0

Il faut, par la suite, effectuer le test d'hypothèse qui nous donne un p-value de 0.6203, toujours plus grand que 0.05. L'hypothèse nulle ne sera donc pas rejetée.

Régression

Comme démontré dans le *heat map* et les autres représentations graphiques présentés dans la section 1, nous remarquons que certaines variables possèdent bel et bien un relation entres elles; par exemple, il est possible de conclure qu'il existe une relation entre les variables metascore et score, et que celle-ci soit pratiquement linéaire. Par ailleurs, le but de cette section s'agit d'affirmer, ou bien de nier, l'existence d'une relation entre deux variables, à l'aide de modèles de régression.

1. Relation entre le nombre de votes et le revenu

Pour cette première relation, nous allons reprendre le graphique présenté dans la section 1, avec l'ajout d'une ligne de régression. Nous avons d'abord essayé avec la régression linéaire.

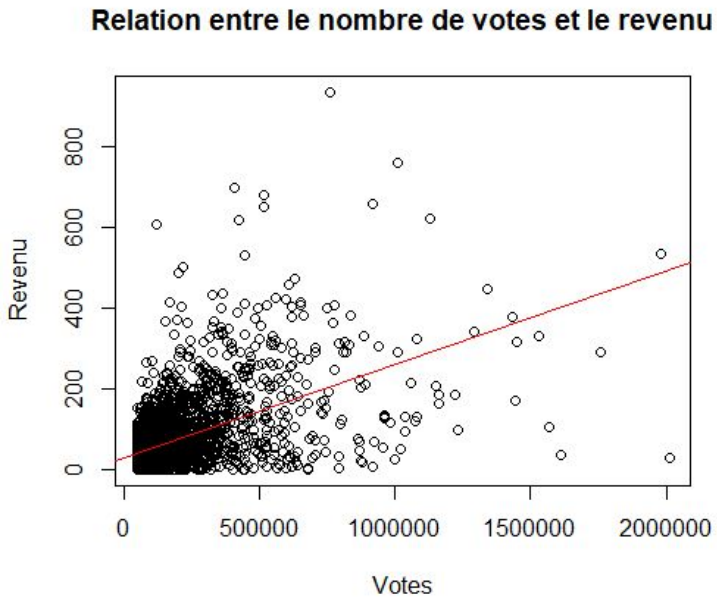


Figure 19. Corrélation entre le nombre de votes et le revenu.

Nous avons vu en classe que lorsqu'un modèle de régression est appliqué, il est préférable que la variance des données soit présentée de façon uniforme. Or, il est facilement discernable que ce n'est pas le cas de ce diagramme, puisque nous retrouvons un amas de points vers le bas de la ligne de régression; plus précisément, on remarque qu'une grande majorité des données se retrouvent dans la zone [0, 500000] x [0, 220] (environ). Déjà là, cela suggère que la dépendance entre ces variables peut être considéré comme étant faible. Puis, si nous nous

fions sur le *heat map* représentant la corrélation de Pearson dans la partie « Statistique descriptive », nous pouvons calculer un coefficient de détermination R^2 équivalant à 0,2704. Ce coefficient est alors très faiblement significatif.

À partir d'ici, nous nous sommes penchés sur la raison pourquoi ce coefficient était aussi basse: est-ce que c'est parce qu'effectivement, ces deux variables sont fortement indépendantes l'une de l'autre, ou bien est-ce que c'est parce que le modèle choisi était inadéquat pour décrire la relation entre celles-ci. Illustrons un diagramme de résidus pour confirmer, ou nier, si le modèle de régression linéaire est adéquate.

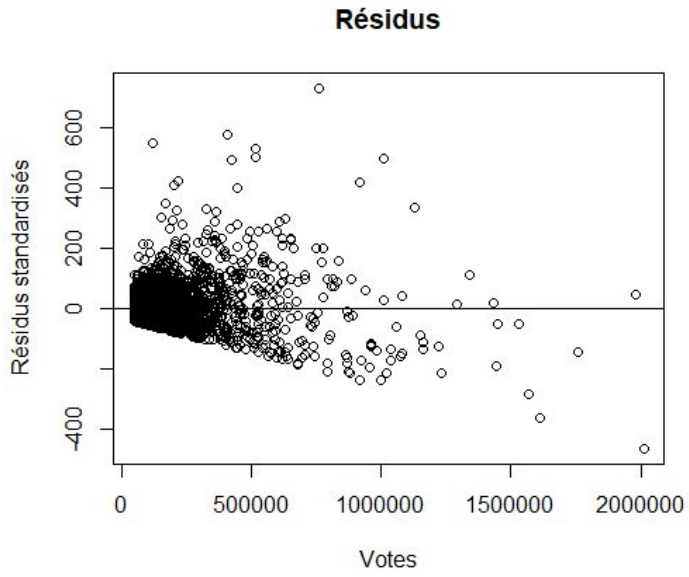


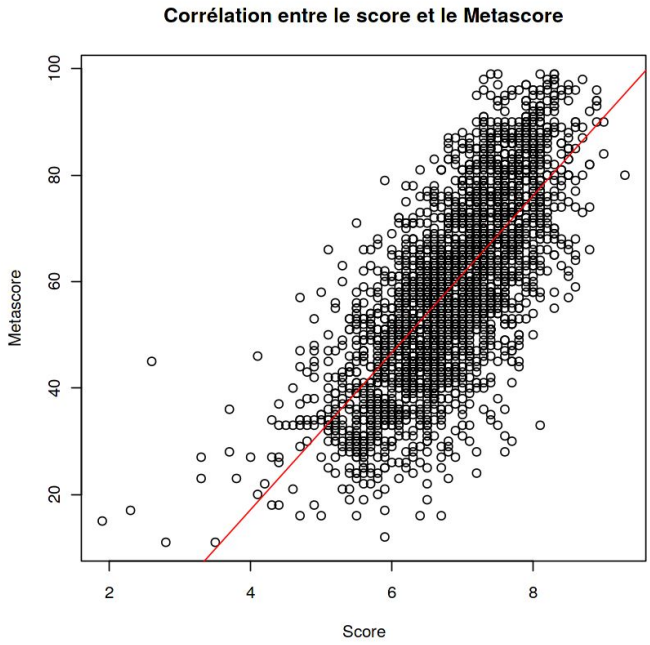
Figure 20. Corrélation entre le nombre de votes et le revenu.

Effectivement, nous remarquons que les points ne sont pas dispersés aléatoirement autour de l'axe horizontal, ce qui suggère qu'un autre modèle serait plutôt approprié [2].

Par contre, étant donné que l'échantillon utilisée pour ce travail est d'une taille assez considérable, il aurait sûrement été préférable de prendre un échantillon d'une taille plus petite à partir de l'échantillon initiale, et ensuite refaire le tout, pour ensuite confirmer si le modèle de régression linéaire est encore inapplicable ou non. Si c'est encore le cas, il aurait été possible d'utiliser un autre autre modèle de régression, soit celui non linéaire par exemple.

2. Relation entre le score du film et le metascore

Encore une fois, pour cette relation, nous avons repris le diagramme utilisé pour cette relation dans la première section.



Comme mentionné au préalable, le coefficient de détermination R^2 de ce modèle est faiblement significatif car il est de 0,5411, ce qui veut dire qu'environ la moitié de la variance observée peut être expliqué par les entrées du modèle.

Conclusion

Selon nos analyses graphiques, le nombre de votes et le revenu accumulé ne sont pas des variables qui sont corrélées de manière linéaire, comme précédemment analysé à l'aide des coefficients de corrélation et de détermination.

Ayant regroupé les genres de films, on peut conclure que, malgré le fait que beaucoup de genres différents ont atteint un très grand nombres durant la même décennie, le groupe 2 est le plus populaire globalement au fils des années, c'est-à-dire les films d'horreur, de crime,

dramatique, mystères et les films noirs. Cela s'explique par sa forte fréquence à chaque décennie.

En produisant des boîtes à moustache pour établir un lien entre le genre de film et le revenu perçu, on observe que le groupe 1 et 3 ont plus tendance à se disperser vers des revenus plus extrêmes, avec des maximums s'approchant de 250 millions de dollars américains. Ce sont aussi ces deux groupes qui ont le plus de données aberrantes ayant des revenus beaucoup plus élevés, allant jusqu'à 600 ou même 800 millions de dollars. On peut en conclure que les groupes 1 et 3, regroupant les genres action, aventure, fantasie, western, sci-fi, famille, animation et comédie, rapportent plus d'argent que les groupes 2 et 4 contenant les genres horreur, crime, film noir, drame, mystère, biographie, musical et romance. On pourrait toutefois améliorer cette analyse en prenant en compte le nombre d'observations de chaque groupe pour uniformiser les groupes.

Avec un coefficient de corrélation de 0.74, la relation entre le score et le Metascore est la plus linéaire parmi toutes les relations possibles. La forme du nuage de points pointe aussi vers la conclusion d'une linéarité.

BIBLIOGRAPHIE

[1] Stack Exchange. (2010) Is normality testing 'essentially useless?. [En ligne]. Disponible: https://stats.stackexchange.com/questions/2492/is-normality-testing-essentially-useless

[2] Delignette-Muller, M.-L., Dutang, C., & Siberchicot, A. *Help to Fit of a Parametric Distribution to Non-Censored or Censored Data*. (2019). PDF.

[3] Stat Trek. Statistics Dictionary: Residual Plots. [En ligne]. Disponible: https://stattrek.com/statistics/dictionary.aspx?definition=residual%20plot

[4] Frost J. Choosing the Correct Type of Regression. [En ligne]. Disponible: https://statisticsbyjim.com/regression/choosing-regression-analysis/