Analyse économétrique du succès économique des films

Florian GOUPIL – Laura CARNAPETE

Introduction

Nous disposons d'un tableau présentant, pour 1665 films sortis notamment en France, les critères suivants : le titre, le nombre d'entrées dans les salles françaises, le nombre de salles dans lesquelles le film est sorti lors de la première semaine, le nombre de titres de presse, média et blogs ayant évalué le film, s'il dispose de documents complémentaires (notamment un trailer) sur Allociné, le nombre de films sortis la même semaine, une mesure de la saisonnalité hebdomadaire du cinéma français sur la période 1993-2009, s'il est un documentaire, un animé, d'origine française ou sorti sur le marché cinématographique américain. Enfin, deux variables de particularité sont associées à chacun d'entre eux.

Dès lors, il s'agit de déterminer les facteurs de succès des films en question (économiques). Afin de répondre à cette problématique, nous allons nous intéresser à cinq variables potentiellement explicatives, et allons les étudier de telle sorte qu'elles mettent en lumière leur relation de corrélation aux profits économiques : le nombre de salles, la médiatisation, Allociné, le cas de particularité 01 et 02, restant à déterminer (qui, a priori, liste les films au succès qui sort des tendances).

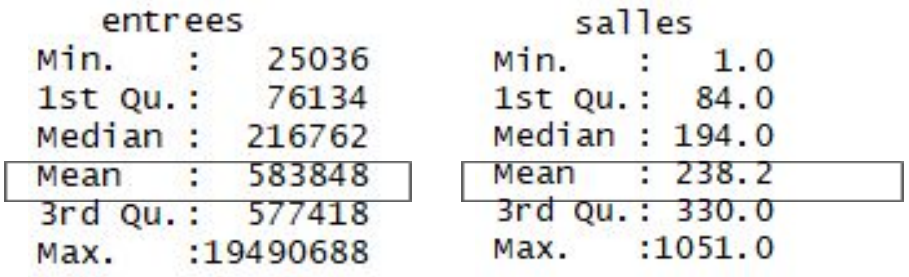
Selon l'hypothèse, qui stipule que ces cinq critères sont ceux qui assurent le plus efficacement le succès d'un film, on en déduit une équation du nombre d'entrées pour un film en fonction de ces facteurs :
y = f (x1, x2 , x3 , x4 , x5)
où
y = nombre d'entrées pour un film
x1 = nombre de salles dans lesquelles le film est sorti lors de la première semaine de diffusion (planification, et donc illustration du budget marketing)
x2 = nombre de titres de presse, média et blogs ayant évalué le film (la couverture médiatique du film)
x3 = oui ou non le film a un support vidéo (bande annonce, trailer, interview…) sur la plateforme française Allociné
x4 = oui ou non le film présente un cas particularité encore inconnu
x5 = oui ou non le film présente un cas particularité encore inconnu

Résumé statistique

Puisque la mission consiste à relever les déterminants du succès des films, à savoir, les curseurs en mesure de « prédire » un quelconque succès, il s'agit de se familiariser avec les données dans leur contexte, et de différencier celles qui « annoncent » plus ou moins efficacement un engouement. Ainsi, dans cette catégorie prédictive, on y trouve, par ordre d'intérêt, par hypothèse : le nombre de salles, puis les autres critères, à part d'intérêt à première vue, égale.

Premièrement, il s'agit de comprendre ce que désignent les deux variables binaires Cas_01 et Cas_02.

Il semble, par conjecture, que la variable Cas-01 recense les films ayant eu bien plus d'entrées que le nombre de salles de diffusion la première semaine laissait entendre, par rapport à la moyenne d'entrées et de salles de diffusion employées à cet effet.
Car, le nombre d'entrées moyen est de 583 848, pour un nombre de salles moyen de 238,2.
De fait, par salle de diffusion ouverte la première semaine, le nombre d'entrées total s'élève autour de 2459, 5. (583 848 / 238,2).



Or, les films mentionnés comme porteurs de la particularité 01 ne respectent pas cette proportionnalité, puisque le nombre d'entrées est très largement supérieur à cette moyenne annoncée.

	titre	entrees	salles
893	Oceanosaures 3D: Voyage au Temps des Dinosau...	425480	1
890	Born to Be Wild	266022	1
343	Under the Sea 3D	262852	1
396	Hubble - Au-delÃ des Ã©toiles	261427	1
1227	Arctique	230852	1
1197	Sugar Man	211428	2
1626	South Pacific	130199	1
1663	Hidden Universe	76111	1
1043	PtÃ©rodactyles 3D : Dans le ciel des dinosaures	72538	1
1633	D-Day, Normandie 1944	67709	1
1571	Dragons 3D: Mythes ou rÃ©alitÃ©	66307	1
742	Tahiti 3D Destination Surf!	33499	1
1321	Adieu au langage	33182	2
1462	20 Feet from Stardom	33039	3
1664	M et le 3Ã¨me secret	30370	2
1599	Jerusalem	27387	1
1123	Week-end	26665	5

En outre, la variable Cas_02 relève les deux films ayant fait le plus d'entrées, si l'on regarde dans l'ordre décroissant des entrées.

	titre	entrees
756	Intouchables	19490688
1562	Qu'est-ce qu'on a fait au Bon Dieu ?	12353181

Voici un résumé statistique des cinq critères retenus, en plus du critère de référence (les entrées) :

Entrées

```
    entrees
Min.    :   25036
1st Qu.:   76134
Median :  216762
Mean   :  583848
3rd Qu.:  577418
Max.   :19490688
```

Nombre de Salles

```
     salles
Min.    :   1.0
1st Qu.:  84.0
Median : 194.0
Mean   : 238.2
3rd Qu.: 330.0
Max.    :1051.0
```

Médias

```
     media
Min.    : 0.00
1st Qu.:13.00
Median :18.00
Mean   :17.65
3rd Qu.:22.00
Max.    :40.00
```

AlloCiné

```
    allocine
Min.    :0.0000
1st Qu.:0.0000
Median :0.0000
Mean   :0.3964
3rd Qu.:1.0000
Max.    :1.0000
```

CAS 01

```
      cas_01
Min.    :0.00000
1st Qu.:0.00000
Median :0.00000
Mean   :0.01021
3rd Qu.:0.00000
Max.    :1.00000
```

CAS 02

```
      cas_02
Min.    :0.000000
1st Qu.:0.000000
Median :0.000000
Mean   :0.001201
3rd Qu.:0.000000
Max.    :1.000000
```

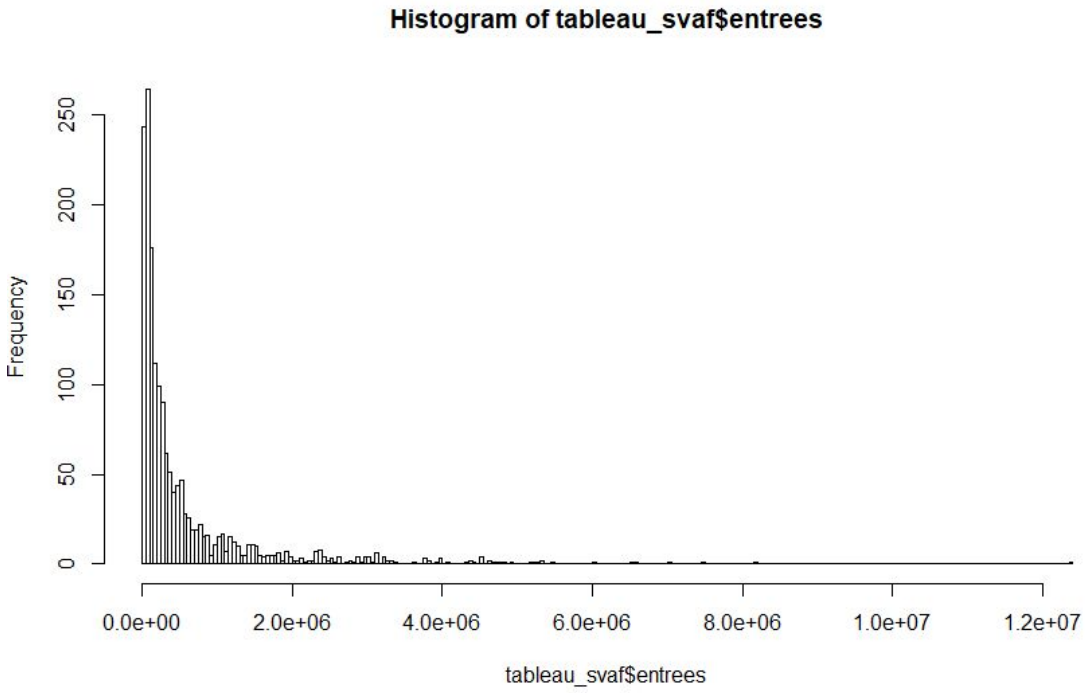
Écarts-types :

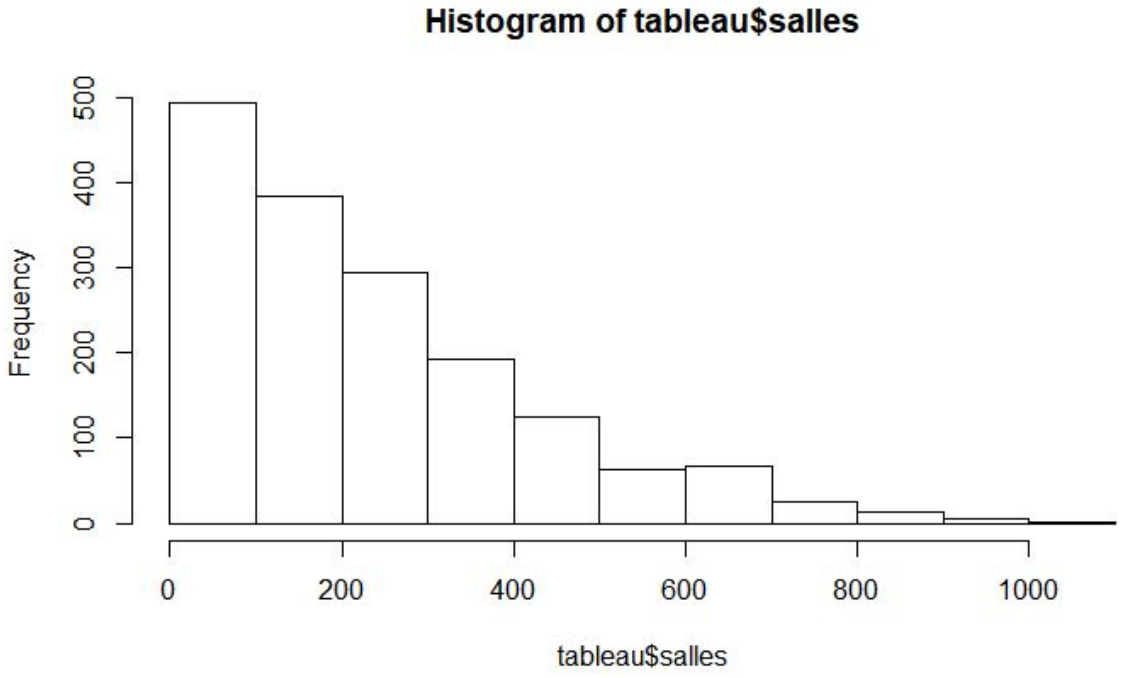
Entrées : 1 073 118

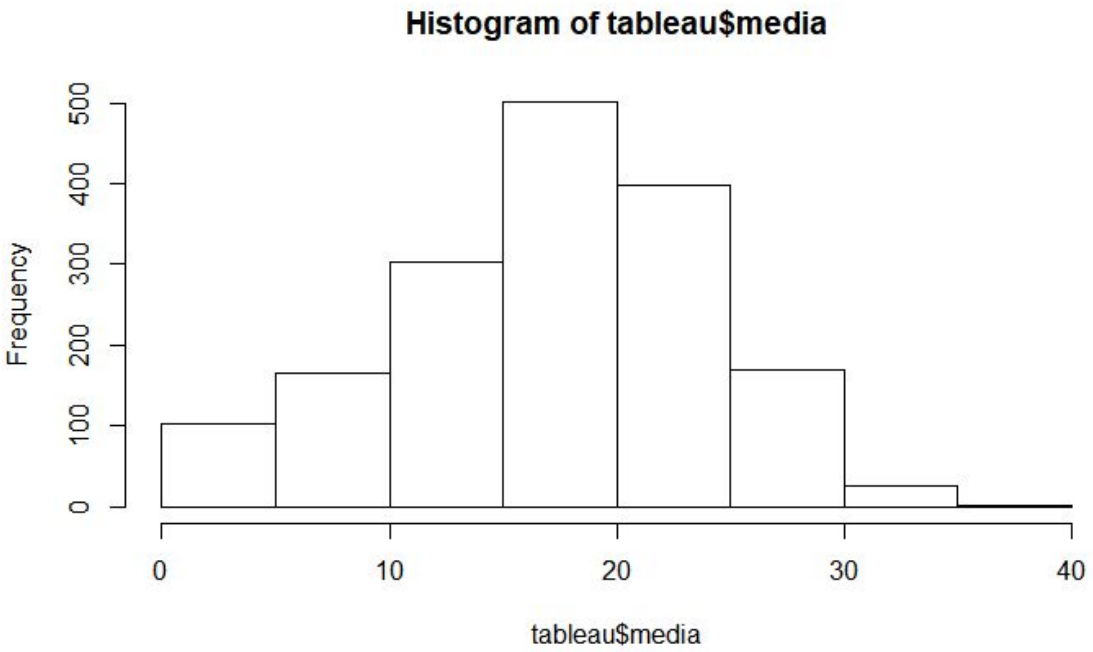
Nombre de Salles : 190,7446	Média : 6,891761

Puis les histogrammes :

Les histogrammes présentés ne sont qu'au nombre de deux : celui des salles, et des médias, en plus de celui des entrées, permettant une référence concrète.







Effectivement, les autres variables sont binaires, pas continues, et de fait, leur représentation par un histogramme ne présente rien de probant, du moins rien qui pourrait nous intéresser, nous aider dans la justification de la pertinence du choix de ces critères.

Ces trois histogrammes nous fournissent plusieurs informations. La dispersion des entrées semble suivre une loi exponentielle, même si des valeurs aberrantes la parasitent. Si les valeurs propres au Cas_02 ont été retirées, le rendu n'en a pas pour autant été affiné. En ce qui concerne le nombre de salles, la dispersion est forte, et les valeurs se concentrent entre 0 et 100 et pas, a priori, de valeurs aberrantes. Sa loi semble être également être exponentielle, et visuellement, une corrélation hypothétique évidente peut-être établie avec les entrées. Du côté des médias, la dispersion est assez faible, les valeurs se concentrent entre 15 et 20 et là aussi, aucune valeur aberrante n'est à relever, et ces valeurs semblent suivre une loi normale. En conséquence, par conjecture, aucun rapprochement évident ne peut être établi à ce stade de l'étude entre le nombre de médias et le nombre d'entrées. La recherche mérite donc d'être approfondie.

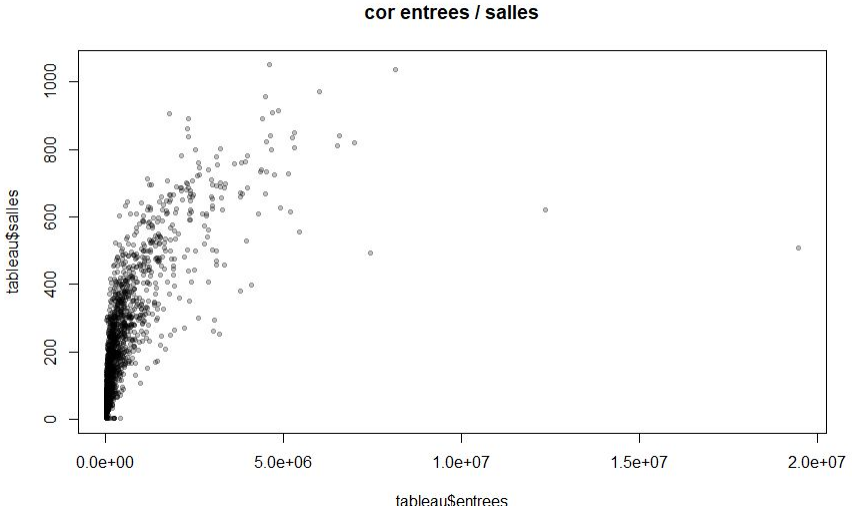
Corrélations linéaires

Voici un récapitulatif des coefficients de corrélations par l'intermédiaire d'une matrice :

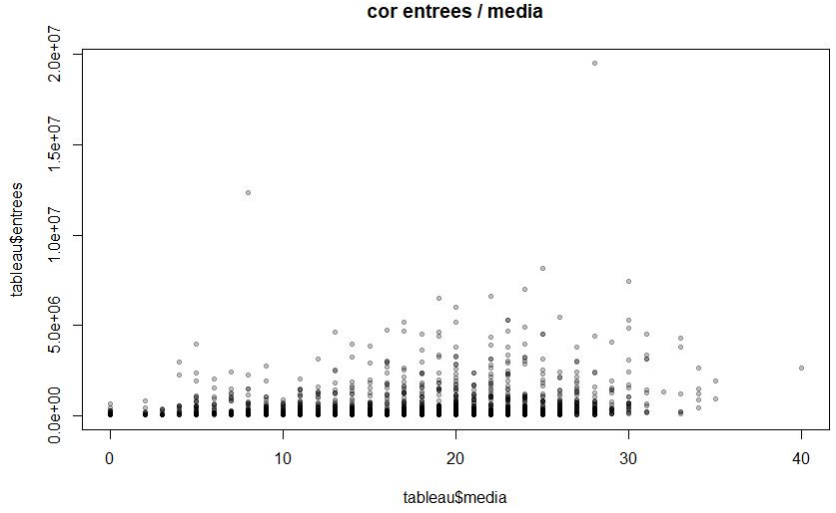
	entrees	salles	media	allocine	autres_films	saison	docu	anim	france	etats_unis	cas_01	cas_02
entrees	1.00	0.70	0.17	0.01	-0.11	0.05	-0.08	0.14	-0.03	0.15	-0.04	0.50
salles	0.70	1.00	0.09	-0.03	-0.11	0.05	-0.19	0.20	-0.04	0.18	-0.13	0.06
media	0.17	0.09	1.00	0.01	-0.06	0.04	-0.15	-0.14	-0.02	0.21	-0.21	0.00
allocine	0.01	-0.03	0.01	1.00	-0.06	0.00	-0.01	-0.05	0.01	0.01	-0.01	0.01
autres_films	-0.11	-0.11	-0.06	-0.06	1.00	0.15	0.08	0.05	0.03	-0.10	0.05	0.01
saison	0.05	0.05	0.04	0.00	0.15	1.00	0.02	0.06	0.00	-0.01	0.01	0.03
docu	-0.08	-0.19	-0.15	-0.01	0.08	0.02	1.00	-0.06	0.04	-0.06	0.41	-0.01
anim	0.14	0.20	-0.14	-0.05	0.05	0.06	-0.06	1.00	-0.08	-0.04	-0.03	-0.01
france	-0.03	-0.04	-0.02	0.01	0.03	0.00	0.04	-0.08	1.00	-0.60	-0.05	0.04
etats_unis	0.15	0.18	0.21	0.01	-0.10	-0.01	-0.06	-0.04	-0.60	1.00	0.01	-0.01
cas_01	-0.04	-0.13	-0.21	-0.01	0.05	0.01	0.41	-0.03	-0.05	0.01	1.00	0.00
cas_02	0.50	0.06	0.00	0.01	0.01	0.03	-0.01	-0.01	0.04	-0.01	0.00	1.00

n= 1665

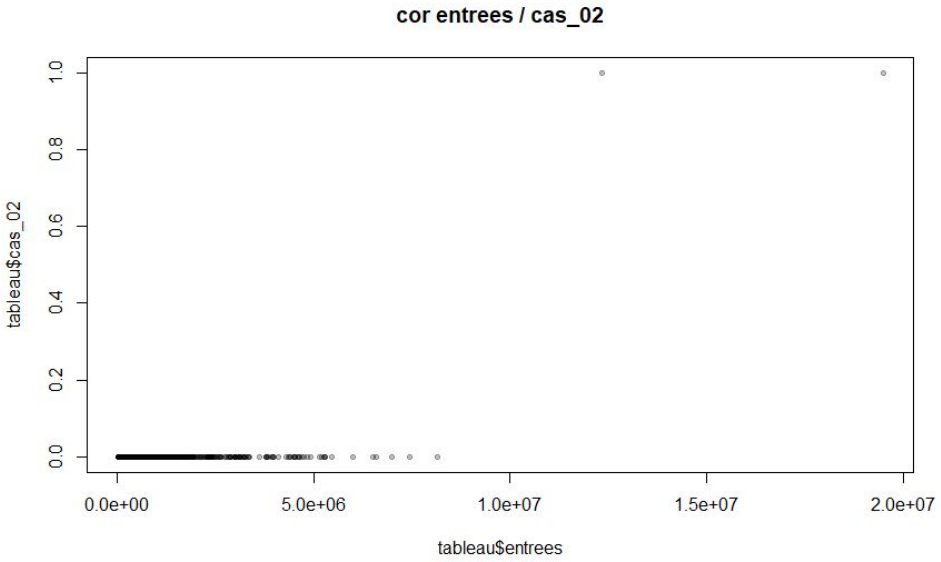
Commençons d'abord par illustrer graphiquement le lien de corrélation entre la variable "patron": les entrées, et les autres variables afin de matérialiser les coefficients de corrélations :



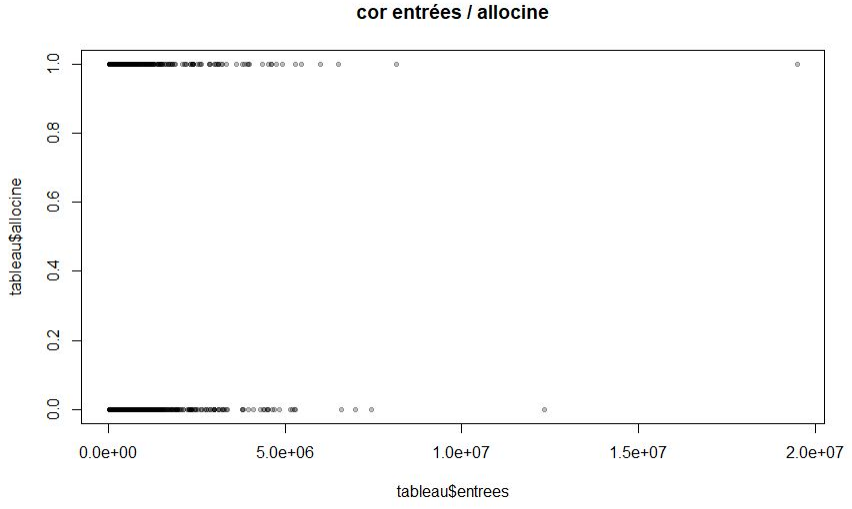
Le graphique semble nous montrer un lien de corrélation positif clair entre l'étendue de diffusion dans les salles et le nombre d'entrées, puisque comme l'indiquait la matrice, le coefficient de corrélation est très élevé (0.7023469 exactement). Cela confirme notre première intuition, déjà appuyée par les histogrammes. Le nombre de salles employées lors de la première semaine paraît tout désigné afin de construire le modèle de régression linéaire multiple pouvant expliquer le succès d'un film.



Ce graphique-ci, visuellement, offre le spectacle d'un lien de corrélation existant, certes, mais peu concluant. Si l'on se réfère à la matrice ci-dessus, le coefficient de corrélation s'élève à 0.17. Ainsi, cela explique l'impression visuelle peu probante concernant la corrélation. Néanmoins, il va sans dire que celle-ci doit être prise en compte car élevée à 17% ; bien plus haute que la plupart des facteurs non-retenus tels que la saison, le type documentaire, l'origine française.

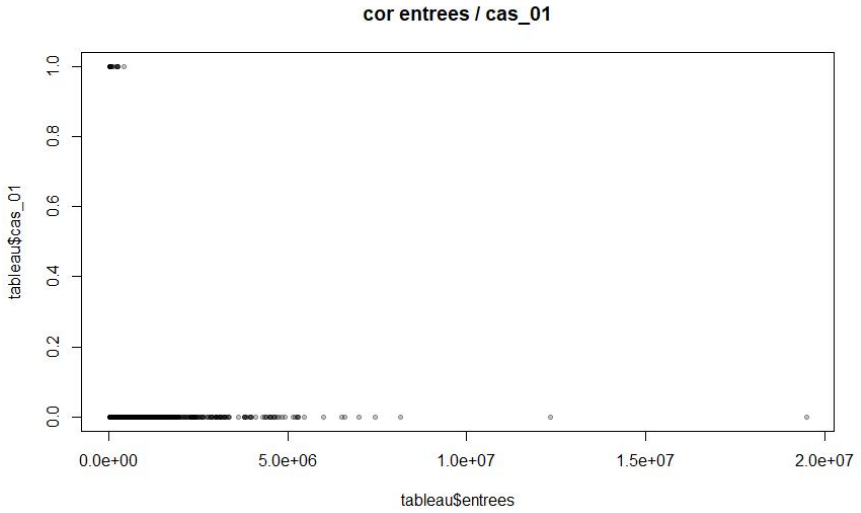


La représentation graphique de corrélation entre les entrées et les Cas_02 est d'aspect, pas concluante. Cependant, certaines précisions méritent d'y trouver place. Premièrement, comme mentionné, la représentativité ne peut pas être considérée comme valide puisque trop restreinte. De fait, la famille de départ (1665 films) devrait être largement agrandie afin que des hypothèses solides puissent être formulées. De surcroît, les valeurs aberrantes que représente le Cas_02 sont bel et bien un événement survenu dans un contexte de normalité, sans pipation et qui concerne les deux films ayant produit le plus d'entrées, laissant donc peu de place au hasard. Ils ne sont donc certainement pas à négliger. Nous décidons par conséquent d'intégrer les Cas_02, représentatifs d'énormément d'entrées (19490688 (entrées Cas_02) / 1665*583848 (entrées totales) = 2%) dans la régression en tant qu'exception.



La corrélation linéaire entre les entrées et le nombre de supports vidéos sur la plateforme Allociné paraît proche de 0. En se référant à la matrice regroupant l'ensemble des coefficients de corrélation, on se rend compte qu'il est à hauteur de 0.01, marquant donc une corrélation très faible. Cependant, le site est incontournable en France, et surtout, marque l'investissement budgétaire en communication des films. Les films concernés par cette caractéristiques représentent énormément de vues et enfin, afin de s'avancer sur la suite de l'étude, on remarque

que c’est l’une des variables présentant le moins de corrélation avec ses consoeures.

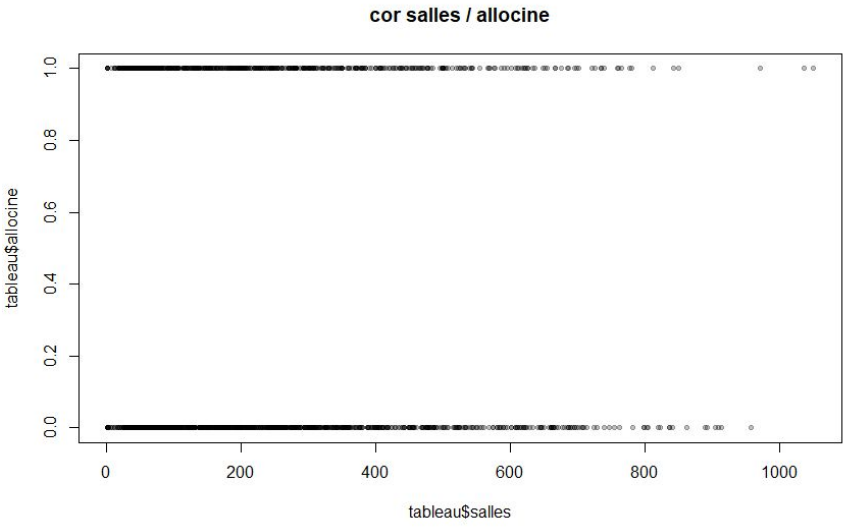


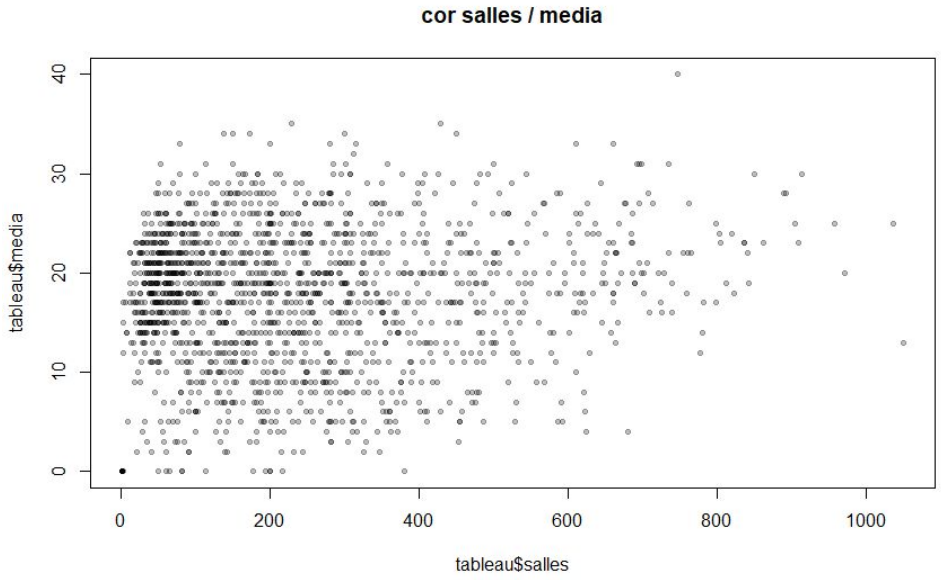
La représentation graphique de corrélation entre les entrées et les Cas_01 est notable, mais pas certaine de cette façon. Comme mentionné pour le Cas_02, la représentativité ne peut pas être considérée comme valide et les valeurs aberrantes qu’ils englobent sont, selon notre point de vue, à prendre en considération comme événement d’intervention à la probabilité non-nulle, et surtout non négligeable. Le Cas_01 rassemble comme le Cas_02 une forte part des entrées, mais cette fois-ci, par rapport à la prédiction générale de succès. C’est pourquoi cette variable sera intégrée dans la régression, en tant que dérive prévisible du modèle.

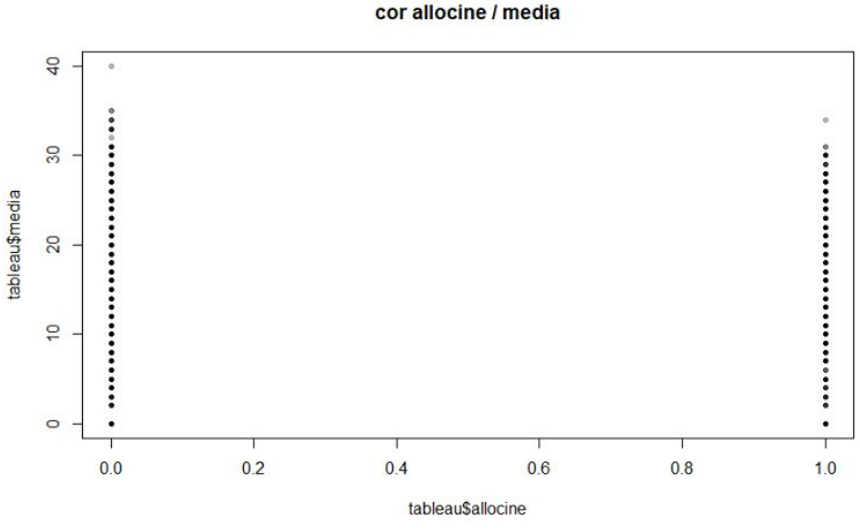
Par ailleurs, avant d’aborder le sujet des corrélations entre les facteurs élus, il est possible d’aplanir le terrain en disant que les coefficients de corrélation tournent de manière globale autour de 0, ce qui consolide notre hypothèse de départ, à savoir que les facteurs salles, média, allocine, cas_01 et cas_02 sont porteurs d’un potentiel “bon”modèle de régression multiple.

Nous avons décidé, en ce qui concerne les critères d’étude, de ne pas représenter par diagramme leur corrélation avec le Cas_01 et le Cas_02. Effectivement, les dix-sept et deux seuls films qu'ils désignent respectivement laissent observer des valeurs d'entrées « aberrantes » car excessivement élevées par rapport au nombre de salle ou élevées de manière générale. De fait, la corrélation, qu'elle existe ou non, ne serait pas justifiable puisque subjective, car fondée sur très peu d'événements, éloignés du seuil d’acceptation d’un échantillon (20).

Il s’agit ici de représenter graphiquement les corrélations linéaires des variables salles, média et allociné. L’intérêt de l’exercice de manière globale est le suivant : s’il advient qu’il existe un véritable lien de corrélation n’ayant pu être détecté précédemment, faute de technique ou de significativité, le modèle de régression linéaire multiple proposé dans ce qui suit ne pourra être “vrai” économétriquement parlant, et donc ne sera pas applicable et généralisable. Notre objectif est donc de lever toutes les incertitudes planant sur ces corrélations parasites.


cor salles / allocine
tableau$allocine
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
tableau$salles


cor salles / media
tableau$media
0
10
20
30
40
0
200
400
600
800
1000
tableau$salles


cor allocine / media
tableau$media
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
tableau$allocine

De manière assez évidente, il n'y a, a priori, aucune objection quant au choix de ces variables, qui, en tout état de cause, ne présentent aucune relation de corrélation susceptible d'être prise en compte et menaçant le modèle en construction.

Modèle de régression linéaire

Comme évoqué précédemment nous avons donc des variables explicatives (Nombre de salles, Media, Allocine, Cas_01, Cas_02) qui présentent, par l'intermédiaire des recherches menées ci-dessus, de bonnes conditions afin de former un modèle valide capable de prévoir le nombres d'entrées réalisées par un film.
Nous avons donc une relation de type y = f (x1,x2,x3,x4,x5). Le modèle inclut également un terme d'erreur ε (epsilon) qui concentrera ce que le modèle ne parvient pas à expliquer et donne la relation de type ε = y - f (x1,x2,x3,x4,x5).
À chaque variable explicative est associé un coefficient que nous évaluons par régression linéaire multiple selon la méthode des MCO, qui est un estimateur sans biais, le plus efficace selon le théorème de Gauss-Markov, sous réserve de validation des hypothèses bien-sûr.

```
Call:
lm(formula = log(entrees) ~ log(salles) + media + allocine +
    cas_01 + cas_02, data = tableau)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1944 -0.4575 -0.0121  0.4711  2.4690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.502203   0.107978  50.957  < 2e-16 ***
log(salles) 1.212137   0.018803  64.464  < 2e-16 ***
media       0.033160   0.002533  13.091  < 2e-16 ***
allocine    0.130591   0.034926   3.739 0.000191 ***
cas_01      5.368716   0.195511  27.460  < 2e-16 ***
cas_02      2.719135   0.493172   5.514 4.07e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6963 on 1659 degrees of freedom
Multiple R-squared:  0.7302,    Adjusted R-squared:  0.7294
F-statistic: 897.9 on 5 and 1659 DF,  p-value: < 2.2e-16
```

1. Analyse des coefficients

Intéressons-nous maintenant aux coefficients.
Nous remarquons que tous sont positifs, ce qui démontre que nos variables expliquent les chances de succès d'un film. À titre d'exemple, pour une augmentation du recours aux salles la première semaine de diffusion de 10%, le nombre total d'entrées incrémente de 12 places environ. On remarque de forts coefficients pour la situation des particularités (Cas_01, Cas_02). Dans cette optique, nous sommes confortés dans le choix d'avoir intégré les valeurs exceptionnelles, puisqu'il semble qu'elles sont absolument déterminantes dans l'exercice.

2. Commentaire du R^2

De plus nous notons que le R^2 ajusté est de 0.7294, ce qui nous permet de soutenir que notre choix de variables est pertinent et qu'elles expliquent en partie le nombre d'entrées de manière satisfaisante puisqu'il respecte le critère imposé ($R^2 > 0.7$). Cela signifie que l'équation de la droite de régression est capable de déterminer environ 72.94 % de la distribution des points.

3. Validation des hypothèses

Maintenant que nous avons montré que la régressions est pertinente, nous devons vérifier que notre régression admet certaines hypothèses :

```
        Shapiro-Wilk normality test

data:  regressioncbase$res
W = 0.9984, p-value = 0.1145

> vif(regressioncbase) # multicolinéarité : cool si <2
log(salles)       media    allocine      cas_01      cas_02
   1.293531    1.046010    1.002426    1.326740    1.002204
> ncvTest(regressioncbase) # homoscedasticité : cool si pvalue >.05
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.599741, Df = 1, p = 0.20594
> shapiro.test(regressioncbase$res) # normalité : cool si pvalue >.05

        Shapiro-Wilk normality test

data:  regressioncbase$res
W = 0.9984, p-value = 0.1145
```

Nous remarquons que les p-values permettent d'affirmer que notre régression respectent les trois hypothèses, puisque dans chaque cas, $p > 0.05 > 0.1$. Ainsi notre modèle économétrique est acceptable et peut désormais être considéré comme un "bon" modèle.

Conclusions

Le modèle nous explique que les salles (budget marketing) sont très impactantes sur le succès économique d'un film. Les autres variables comme le référencement sur Allociné, ou le nombre de médias employés à sa critique influent relativement peu sur le succès d'un film. Enfin, les variables répertoriant les valeurs aberrantes du nombre d'entrées en fonction des paramètres de départ (nombre de salle …), et plus simplement de l'attente médiatique, sont absolument à inclure à la modélisation si celle-ci doit-être "fiable", et ce, malgré les difficultés d'étude de corrélation qui les accompagnent. Car un "bon" modèle doit prévoir sa propre marge d'erreur.

Puisque le modèle est abouti, nous entreprenons de le tester sur un exemple concret d'application pour observer son emploi dans un autre contexte. De cette manière, nous allons pouvoir évaluer son efficacité "sur le terrain" en plus de son efficacité théorique prouvée.

	titre	entrees	salles	media	allocine	autres_films	saison	docu	anim	france	etats_unis	cas_01	cas_02
482	Buried	226838	228	23	1	7	0.271033110	0	0	0	1	0	0

Calcul d’estimation :

5.502203+1.212137*log(228)+0.033160*23+0.130591*1+5.368716*0 = 12.97658 (entrées prévues)

Valeur effective :
log(226838) = 12.3319

Calcul de l’écart relatif :
(12.97658 - 12.3319) / 12.3319 = environ 5.23%
Ainsi, le modèle présente une faible marge d’erreur, comme déduit précédemment.