**MODELISATION STATISTIQUE**

*Introduction*

La modélisation statistique est une manière mathématique d’approximer la réalité. Autrement dit, elle permet d’expliquer les processus qui créent les données. En plus de cela, elle permet également de réaliser des prédictions à partir de ces approximations.

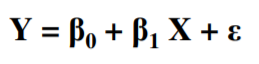
Dans un modèle statistique, il y a, dans la majeure partie des cas, « une variable dépendante » et une ou plusieurs « variables indépendantes ». Aussi appelées « variable à expliquer » et « variable explicative », qui ont plus de sens.

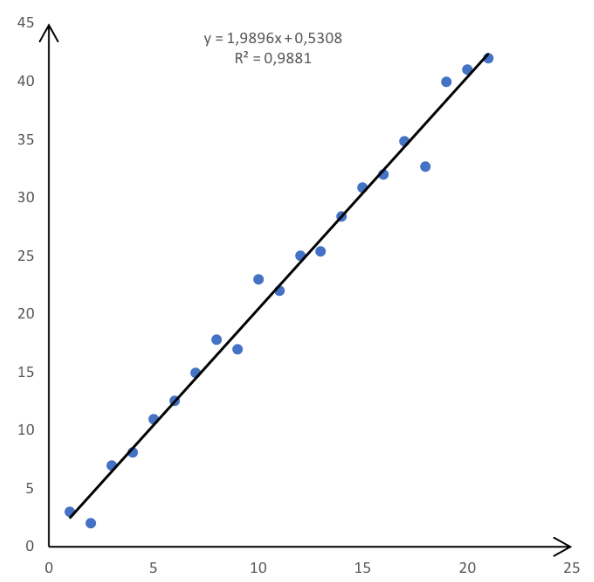
En effet, grâces à nos modèles nous allons pouvoir expliquer quelles « variables indépendantes » prédisent la variable « dépendante ».

Plusieurs modèles existent : Régression linéaire simple, régression linéaire multiple, régression logistique, …

*Les modèles*

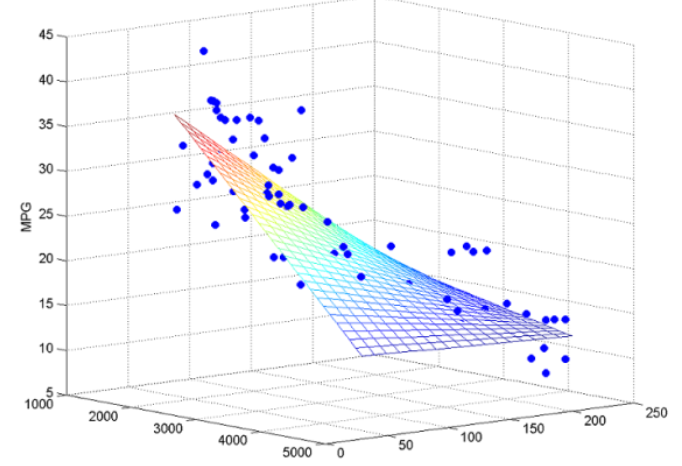
**La régression linéaire simple** est un modèle statistique utilisé lorsque que nous avons une seule variable indépendante et nous voulons savoir si elle prédit ou non notre variable dépendante.



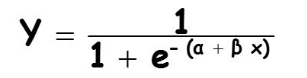
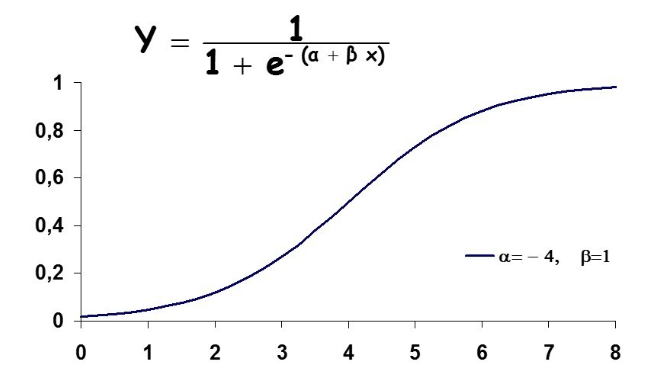
****

**La régression** **linéaire multiple** est un modèle statistique utilisé lorsque nous avons plusieurs variables indépendantes et nous voulons savoir si elles prédisent ou non notre variable dépendante.





**La régression logistique** est un modèle statistique dans lequel notre variable dépendante sera une variable booléenne. C’est-à-dire qu’elle ne peut prendre comme valeur que 0 ou 1 (Non ou oui). Ce modèle va permettre, via la variable indépendante, de prédire la probabilité qu’un élément va se produire ou non.



*Choix du logiciel*

Les choix se présentant à nous était large, nous avons néanmoins voulu intégrer des connaissances acquises par nous même au sein du projet plutôt que d’utiliser R, déjà utilisé dans d’autres parties. Nous avons donc choisi de nous orienter vers GRETL.

GRETL est un logiciel, « open source » et programmé en C, de statistiques. Sa première version est sortie en 2000. Il peut être utilisé autant en ligne de commandes via le langage « hansl » qu’en interfaces graphiques. Le logiciel paru plusieurs fois dans les journaux de sciences tel que dans le ‘’Journal of Statistical Software’’ en 2008, vendant ses mérites. Sa communauté est forte, il possède même une conférence se produisant tous les 2 ans depuis 2009.

La raison de ce choix est que GRETL est très intuitif, facile à utiliser et performant.

*Analyse*

Nos datasets, GAME et GLOBAL, sont composé de 11 colonnes chacun. Dans ces colonnes, aucune ne comporte de valeur booléenne. Donc nous pouvons déjà en déduire que le modèle à régression logistique est à éliminer. Il nous reste donc nos deux modèles de régression linéaire. De toute évidence, vu le nombre de colonnes que nous avons, un système de régression linéaire simple n’est pas envisageable.

Nous voilà donc avec notre système de régression linéaire multiple. Nous allons devoir prendre en compte que nous avons des valeurs catégoriques pour « CATEGORY » et « CONTENT\_RATING ». Ces colonnes devront être modifiées dans GRETL pour stipuler qu’il s’agit de variables discrètes.

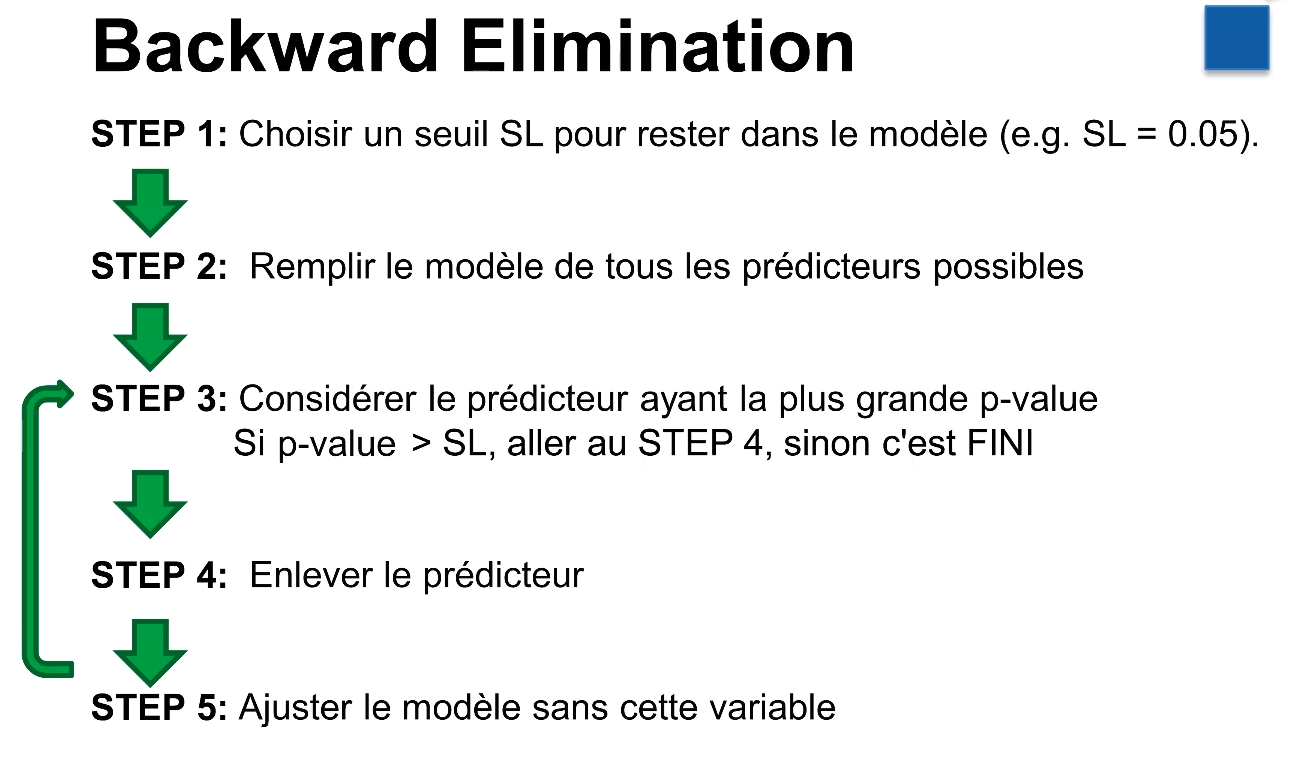
Nous allons donc prendre comme variable dépendante le nombre d’installations et comme variables indépendantes, toutes les autres. Nos variables discrètes devront être créées dans GRETL via la fonction « variables discrètes ».

*Choix de la méthode de construction du modèle*

Il existe plusieurs méthodes de construction de modèles ; All-in, Backward-elimination, Bidirectional Elimination, …

Chacune ont leurs points forts et leurs points faibles, malheureusement nous ne connaissons qu’une méthode de construction qui est la « BACKWARD-ELIMINTION ». Nous allons donc utiliser cette méthode.

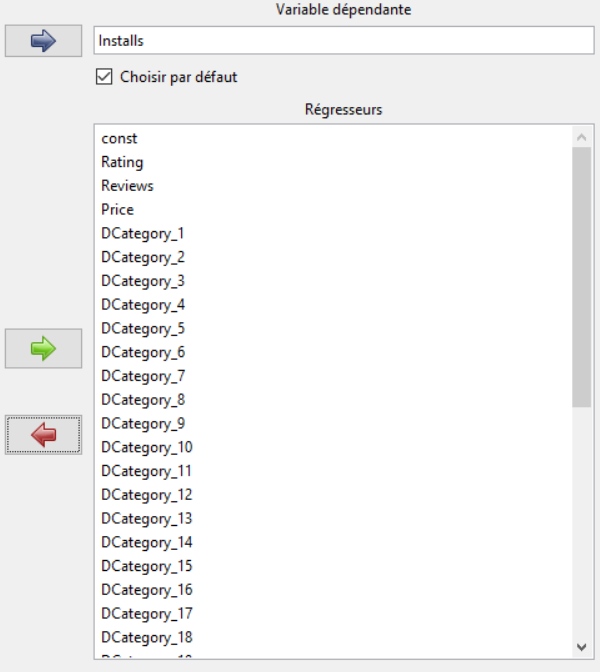
Comment fonctionne-t-elle ?



*Mise en pratique – Dataset GLOBAL*

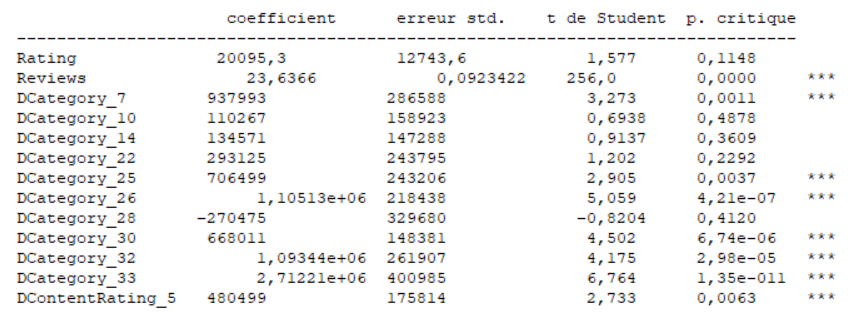
Après avoir choisi comme seuil 5% et spécifier le modèle dans GRETL, il nous suffit de déclarer la variable dépendante (Installs) et les variables indépendantes. En ce qui concerne les « variables discrètes », il faut respecter la règle de ne prendre que les n-1. Dans notre cas nous avons 34 variables discrètes de CATEGORY, nous n’en prenons que 33. Dans content nous avons 6 variables discrètes, nous n’en prenons que 5.

Les paramètres dans GRETL devraient ressembler à ceci :

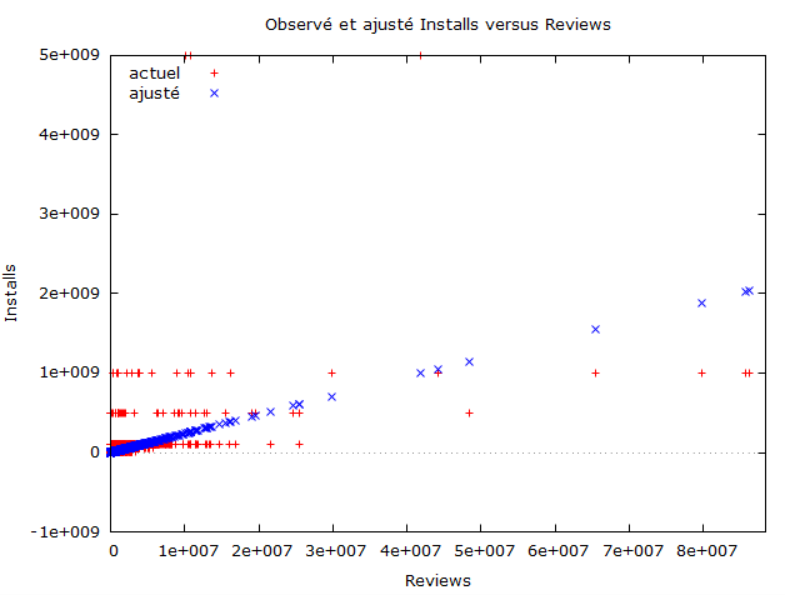


Une fois les répresseurs, nous pouvons valider notre modèle. Nous obtenons, via GRETL, une fenêtre indiquant les p-values des variables indépendantes. Maintenant il ne nous reste plus qu’à appliquer le modèle « BACKWARD-ELIMINATION ». Il nous faut donc éliminer la plus grande p-value et refaire le modèle jusqu’à ce que toutes les p-value respecte le seuil de 5%.

Une fois fini, la méthode du BACKWARD ELIMINATION nous donne ceci :

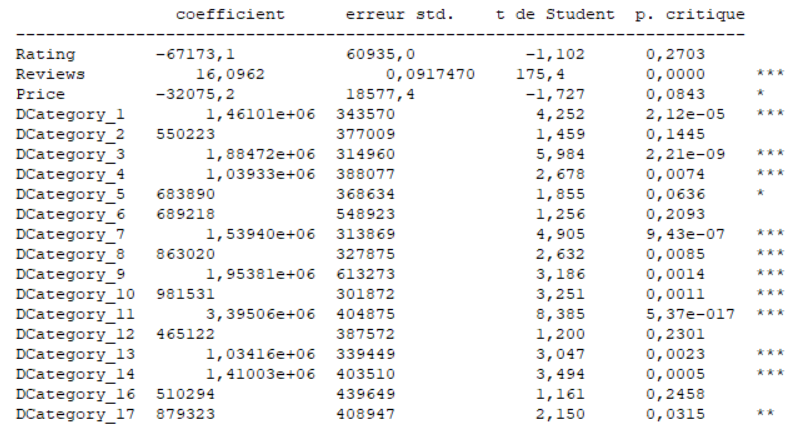


Qu’est-ce que ceci signifie ? Notre p-critique tout à droite est ce, sur quoi nous nous sommes basés pour notre méthode. En effet, nous pouvons remarquer qu’il n’y a plus de p-value supérieur à 5%. De plus, GRETL nous spécifie via un système d’étoile, si une variable indépendante est très prédictive ou non. Ici nous pouvons surtout remarquer que la p-value de Reviews est de 0. Nous avons donc ci-dessus, les variables indépendantes qui prédisent au mieux notre variable dépendante « Installs ».



*Mise en pratique – Dataset GAME*

Sans grand étonnement, nous allons ici procéder de la même façon que précédemment. Nous allons réutiliser la méthode du « BACKWARD – ELIMINATION ». Nous prenons à nouveau un seuil de 5 % et nous réutilisons les mêmes variables indépendantes.



Nous pouvons constater que nous avons beaucoup de variables indépendantes qui peuvent prédire notre variable dépendante. A nouveau nous pouvons constater que la variable le prédisant le mieux est Reviews qui a à nouveau obtenu une p-value de 0.

