Modélisation statique avancée
Deuxième année STID – IUT
Université de Paris
Enseignant : Arthur Leroy
Email : arthur.leroy.pro@gmail.com

# Régression linéaire multiple, points influents et atypiques

## Déroulement :

Cette partie se déroulera en 3 séances :

- 1ère séance : Formation des groupes + Mise en place du travail attendu + Eléments de cours, recherche de sources supplémentaires.
- 2ème et 3ème séance : étude statistique en *autonomie dirigée* par groupes.

## Travail attendu :

Le but de ce module est de mettre en application ce que vous avez appris lors des Cours/TP/TD de Modèle Linéaire sur un jeu de données réelles. Lors de la première séance vous commencerez par former des groupes de 3 étudiants. A la suite de quoi vous mettrez en place le plan de votre étude à venir. Lors des 2ème et 3ème séances vous réaliserez cette étude à l'aide du logiciel R. Enfin vous me rendrez un rapport détaillé sur le sujet. J'attends de vous une mise en application rigoureuse des méthodes présentées dans vos différents cours. De plus, une prise d'initiative est attendue, en cherchant par soi même des informations, du contenu, des packages, sur internet par exemple. Cet aspect sera mis en valeur dans la note finale.

*Vous serez bientôt en stage, où il vous sera demandé une certaine autonomie. Le format de ce cours a pour objectif de vous aider à acquérir cette autonomie du mieux possible tout en conservant un support pour vos questions et vos progrès, profitez en ! Je vous laisserai un maximum travailler par vous même en groupe, mais reste disponible dès que vous aurez besoin d'aide, n'hésitez pas, même lorsque nous serons en distanciel.*

## Consignes d'évaluation :

- Vous me rendrez votre rapport sur **Moodle au format PDF ou HTML exclusivement.**
- Vous le deposerez au plus tard le **29 janvier 2021 à 23h55.**
- **Attention :** Moodle est programmé de sorte qu'il ne vous sera plus possible de transmettre votre rapport au-delà de cette date.
- La rédaction et la présentation de votre rapport seront prises en compte dans la note.

## But de l'étude de cas :

On s'intéresse à un échantillon de données réelles sur des habitations résidentielles provenant de Wake County en Caroline du Nord USA. Wake County est l'endroit où se trouve Raleigh, la capitale de l'état de Caroline du Nord mais aussi Cary. Ces deux villes ont respectivement la quinzième et huitième plus grosse croissance de ville aux Etats-Unis faisant de Wake County le neuvième comté ayant la plus grosse croissance du pays. Wake County a connu une hausse de 31.18% de croissance dans sa population depuis 2000, avec une population d'approximativement 823345 personnes.

Le but de cette étude de cas est de voir quelles sont les variables qui influencent le plus le prix des résidences à Wake County.

On effectuera une selection de variables afin de déterminer celles qui jouent un rôle important et on ne conservera que ces dernières dans le modèle final "simple". Il vous faudra regarder attentivement les points influents et/ou atypiques mais également s'intéresser à la validation de votre modèle. Vous préciserez les points qui semblent poser problème. Pour cela vous préciserez les points en question, le critère et si vous les retirez ou non de vos données. Veillez à bien justifier et détailler cette partie.

Enfin vous conclurez votre rapport en donnant deux conclusions à cette étude :

1. La première s'adressant à un spécialiste de la statistique.
2. La seconde étant destinée à un non spécialiste de la statistique.

## Jeu de données :

Le jeu de données à étudier est *wakecounty* qui se trouve sur Moodle au format csv.
Il contient des statistiques sur $n = 100$ habitations. Le jeu de données se compose des 11 variables suivantes :

— *ID* : numéro de l'habitation.
— *Total* : la valeur totale de la propriété en dollars.
— *YearBuilt* : l'année de construction de l'habitation.
— *SqFt* : la surface de l'habitation en feet au carré.
— *Story* : le nombre d'étages de l'habitation.
— *Acres* : le nombre de demi-hectare.
— *NoBaths* : le nombre de salles de bain dans l'habitation.
— *Fireplaces* : le nombre de cheminées dans l'habitation.
— *Land* : la valeur du terrain en dollars.
— *Building* : la valeur du bâtiment de l'habitation en dollars.
— *Zip* : le code zip de la propriété.

## Travail préliminaire sur les variables :

— On commencera par faire des statistiques descriptives afin de rechercher les valeurs manquantes que l'on traitera par suppression des individus.

— Dans notre modèle linéaire on ne prendra pas en compte les variables : *ID* et *Zip* car elles servent seulement à identifier les habitations.

— On ne prendra pas en compte non plus la variable *Land*. Justifier pourquoi.

$\Longrightarrow$ On cherchera donc à expliquer la variable *Total* à l'aide des 7 autres.

## Quelques règles de rédaction d'un rapport :

Le rapport est un document qui doit être de qualité avec une présentation réfléchie et structurée. Il a pour but de présenter le cheminement de votre étude. Ce n'est pas une liste exhaustive des travaux effectués mais une vitrine du travail effectué. Voici quelques règles simples à respecter.

### Organisation du rapport

— Le rapport commence toujours par une introduction générale du sujet d'intérêt.

— Ensuite vous présenterez la base de données : contexte et but de l'étude, présentation des différentes variables, problématique.

— Viennent ensuite les résultats pertinents de statistiques descriptives. Les statistiques descriptives peuvent être résumées sous forme de tableaux et de figures et doivent être commentées (valeurs pertinentes pour l'étude, valeurs atypiques, valeurs manquantes, dissymétrie des distributions, ...).

— L'étude des relations entre les variables doit être présentée avec soin. Pour chaque relation étudiée, présentez précisément les variables étudiées, l'outil statistique utilisé, le résultat de la comparaison ou du test, la conclusion que l'on peut en tirer. Terminez par un commentaire replaçant le résultat obtenu dans le contexte de l'étude.

— N'oubliez pas d'interpréter **TOUS** vos résultats.

— Vous conclurez votre rapport en donnant deux conclusions à cette étude :

1. La première s'adressant à un spécialiste de la statistique.
2. La seconde étant destinée à un non spécialiste de la statistique.

— Si vous insérez des annexes veillez à ce qu'elles soient en nombre limité.

## Références :

[1] Le polycopié de Philippe Besse *"Pratique de la modélisation statistique"* et son site http ://www.math.univ-toulouse.fr/ besse/enseignement.html

[2] Azais, J. M., & Bardet, J. M. (2006). Le modèle linéaire par l'exemple : régression, analyse de la variance et plans d'expériences illustrés avec R, SAS et Splus. Dunod.

[3] Cornillon, P. A., Guyader, A., Husson, F., *et al.* (2008). Statistique avec R.

Vous pouvez trouver d'autres références dans l'introduction de votre cours sur le modèle linéaire.