ANALYSE STATISTIQUE & LANGAGE R

Tests d’hypothèses

Déroulement

- Principes des tests de dépendances / indépendances de variables
  - *Test du khi2 / chi-square / X^2*
  - *Test d'Anova (analyse de variance)*
  - *Exercice pratique*
- Modèle de régression linéaire
  - *Méthode des moindres carrés, coefficient de détermination*
  - *Splitter le jeu de données en train / test*
  - *Entrainement du modèle (2 variables, tracer la régression)*
  - *Tester la précision*
- Evaluation Notebook R
  - *Tracer des graphiques d'analyses descriptives*
  - *Tests de dépendances / indépendances*
  - *Créer un modèle de régression linéaire*
  - *Savoir interpréter les résultats*

Pourquoi tester l’indépendance des variables?

- Eviter la multicolinéarité
- Ne pas avoir des variables qui mesurent la même chose
- Instabilité des coefficients
- Difficulté d’interprétation / explicabilité du modèle

Les tests d’indépendances

- Variables quantitative – quantitative : **Matrice de corrélation**
- Variables qualitative – qualitative : **Khi2**
- Variables quantitative – qualitative : **Anova**

Loi du Khi2

- Tests d'adéquations / ajustements
  - *Comparer deux distributions / deux séries*
  - *Tests d'indépendances*

- Deux risques d'erreur
  - *Risque de 1er espèce / de type I : rejet de H0 l'hypothèse nulle*
  - *Risque de 2^{e} espèce / de type II : non rejet de H0 l'hypothèse alternative*

H0 : le médicament n'a pas d'effet / H1 : le médicament a un effet

Exemple de test Khi2

- Réaliser un tableau de contingence

22 => 27 * 0,83

5 => 27 – 22

	Page Assurance	*Accueil*	*Espace Particulier*	***Total***
Desktop	34 **(35)**	2 **(22)**	154 **(133)**	190 **(83%)**
Mobile	8 **(7)**	25 **(5)**	6 **(27)**	39 **(17%)**
Total	42	27	160	229

- Calcul du Khi2

$$X^2 = \frac{(34-35)^2}{35} + \frac{(2-22)^2}{22} + \frac{(154-133)^2}{133} + \frac{(8-7)^2}{7} + \frac{(25-5)^2}{5} + \frac{(6-27)^2}{27} = \mathbf{118}$$

Exemple du khi2

- Nombre de degrés de libertés (ddl)

Nombre de v.a (non déterminé par une équation)

Nombre de colonnes – 1 * nombre de lignes – 1

(2-1) * (3-1)

- Table de la loi du khi2

Seuil de signification α = 5%

Degrés de liberté	Valeurs du χ^2										
α / Pvaleur (probabilité)	0,95	0,9	0,8	0,7	0,5	0,3	0,2	0,1	0,05	0,01	0,001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82

Mesurer l’intensité de la relation

- Quantifier le lien qui existe entre ces deux variables qualitatives
- V Cramer
  - *Permet de mettre le résultat sur une échelle [0 ; 1]*

X^2 : la valeur du khi2

X^2 max : effectif * [min (nb de lignes ou nb colonnes) - 1]

$$V = \sqrt{\frac{X^2}{X^2 max}}$$

$$\sqrt{\frac{118}{229*1}} = 0.72$$

Exercice Khi2

- Jeu de données : https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

- Tableau de contingence : table(x, y)
- Valeur du khi2 : chisq.test(table)
- *Retranscrire la formule de V de Cramer*
- Réaliser le Khi2 pour :
  - *education + job*
  - *housing + loan*

Anova : Analysis of variance

- On cherche à expliquer la variance inter classes sur la variance intra classes
- Déterminer si les valeurs de la variable quantitative s'organisent selon les modalités de la variable qualitative
- On va donc comparer la moyenne au sein des groupes et entre les groupes

- Exemple :

loyers	650	700	620	
voiture	85	98		
alimentation	160	180	140	120

Anova : formules

- Même cadre contextuel que la régression linéaire
- **Somme des écarts intra classe**

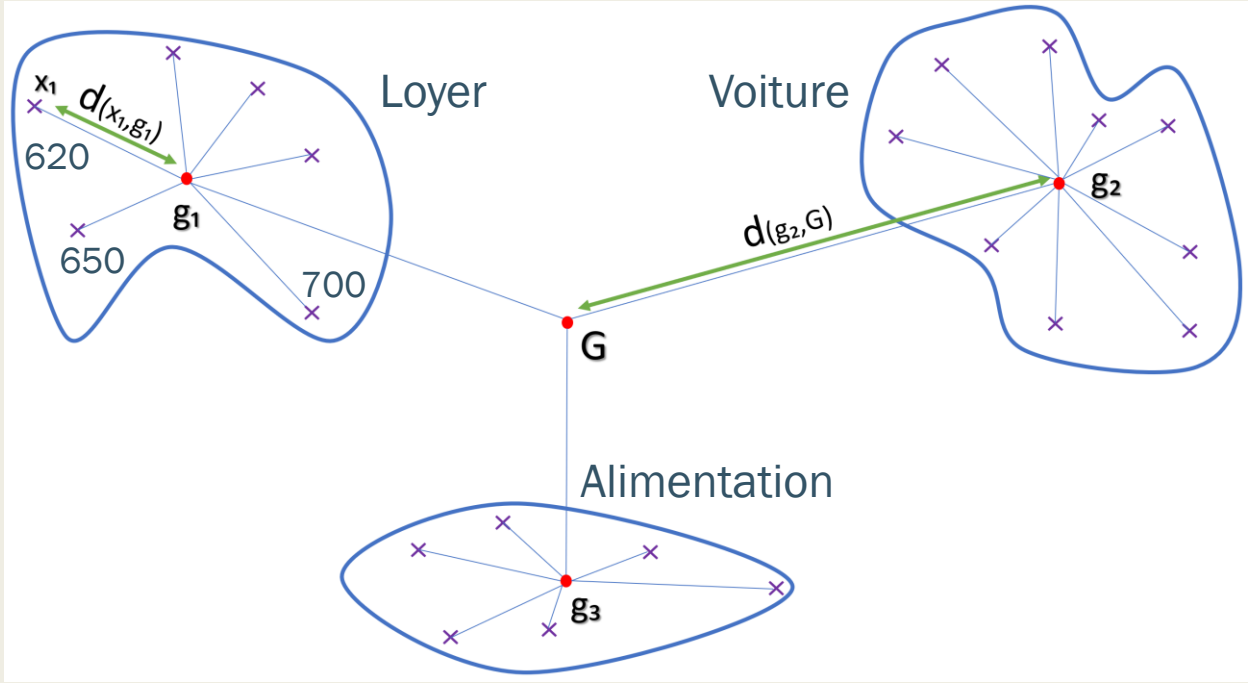
$$\sum_{i=1}^{k}\sum_{j=1}^{n}(yij - \overline{yi})^2$$

- **Somme des écarts inter classe**

$$\sum_{i=1}^{k}(\overline{yi} - \overline{y})^2$$

- **Somme des carrés totale**

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(yij - \overline{y})^2$$



Anova : résultat

- $F\ Value = \dfrac{\dfrac{Inter}{K-1}}{\dfrac{Intra}{N-K}}$

N : les effectifs
K : les classes

La F Value donne une statistique pour la loi de Fisher, à reporter dans une table de loi de Fisher (α=5%, α=10%...), avec autant de table que de seuil de signification. La statistique est obtenue en reportant le nombre de ddl Intra et Inter classes.

ddl intra : *nb classes * (nb de valeurs - 1)*

ddl inter : *nb classes - 1*

La variation inter classes doit être supérieure à la variation intra classes

Anova avec R

```
            Df Sum Sq Mean Sq F value   Pr(>F)
group        2 558266  279133     313 8.56e-07 ***
Residuals    6   5351     892
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La F-Value est ici de 313, avec un indice de confiance élevé : **P-Value faible**

Ce qui signifie qu’au sein de chaque classe les moyennes sont homogènes, mais très hétérogènes entre les groupes

Une F-Value approchant 0 indique que les moyennes entre les groupes sont proches de la moyenne générale, donc pas de distinction

Anova facteurs

- Analyse à un facteur : 1 variable qualitative
- Analyse à deux facteurs : 2 variables qualitatives
- Analyse multifactorielles : X variables qualitatives
- **Toujours 1 variable quantitative et 1 à X variables qualitatives**

Si on possède une autre variable quantitative, possibilité de la transformer :

90 => 90cv

120 => 120cv

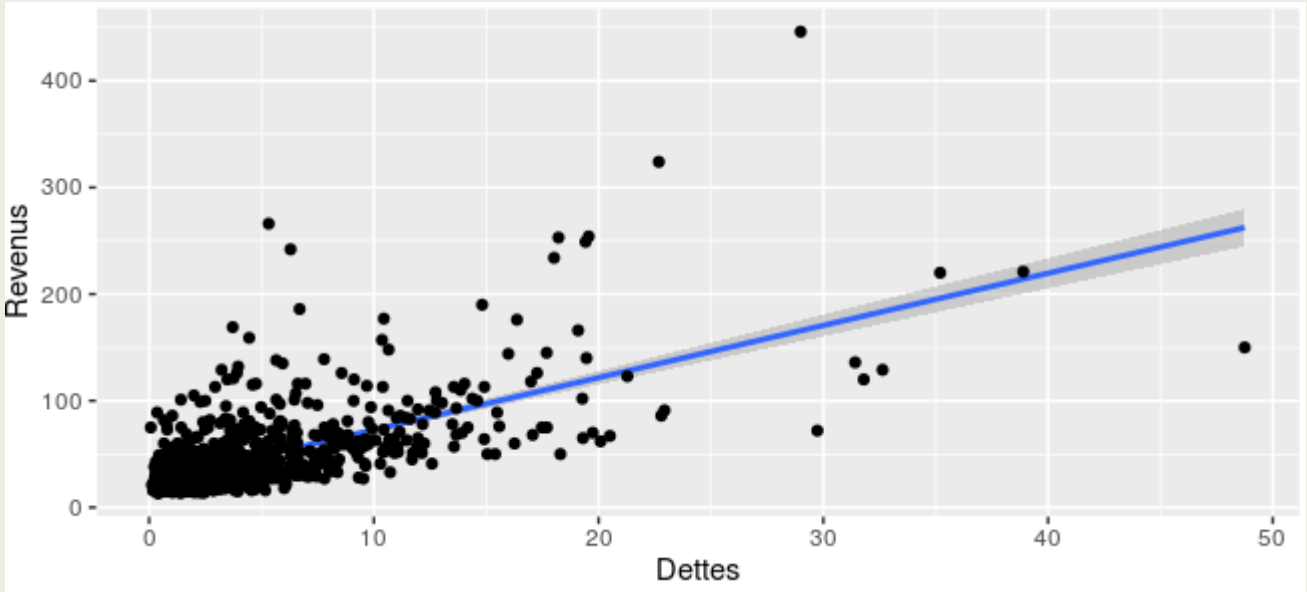
140 => 140cv

Exercice Anova

- Fonction aov(y ~ x + z)
- Réaliser plusieurs Anova
  - *Anova à 1 facteur : age ~ marital*
  - *Anova à 2 facteurs : age ~ marital + loan*

La régression linéaire

- Fonction affine de type ax+b
- Chercher à détecter une relation linéaire entre une variable à expliquer (revenus), et une à N variables explicatives (dettes)
- L'ajustement de la droite se fait via la méthode des moindres carrés
- Problème de régression / ~~classification~~ (régression logistique)



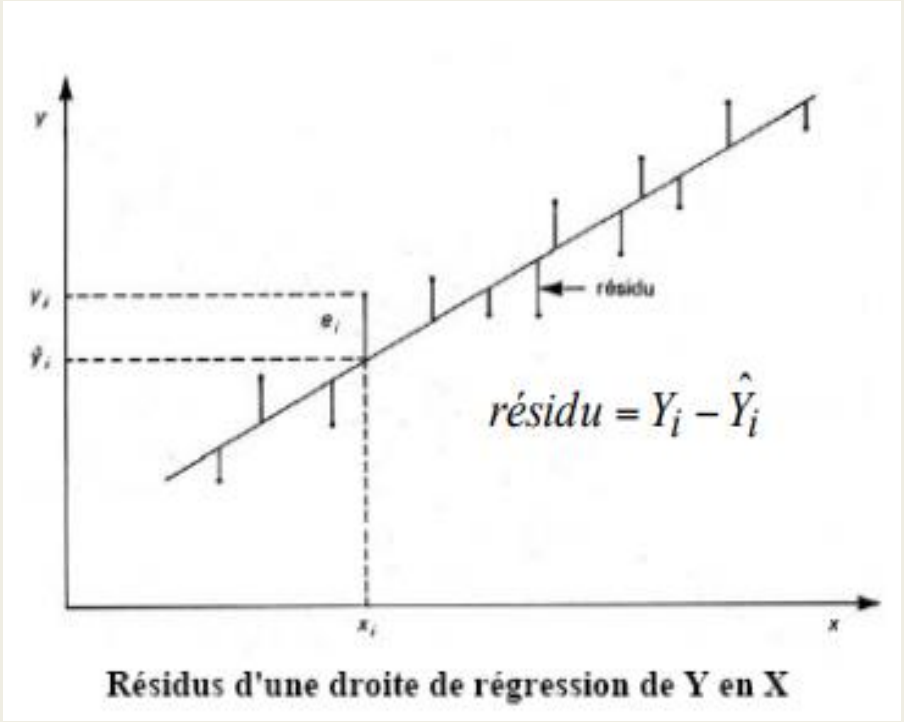
Méthodes des moindres carrés

- MSE (Mean Squared Error) :

$$\frac{\sum(yi - \hat{y}i)^2}{N}$$

- Moyenne des écarts entre chaque point et l'équation de la droite, au carré
- Ajuster la droite pour approximer au mieux le nuage de points
- RMSE (Root Mean Squared Error) : valeur absolue

$$\sqrt{\frac{\sum(yi - \hat{y}i)^2}{N}}$$



Résidus d'une droite de régression de Y en X

Coefficient de détermination

- Evaluer la qualité d'une régression linéaire
- R2 :

$$\mathbf{1} - \frac{\sum(\boldsymbol{yi} - \widehat{\boldsymbol{y}}\boldsymbol{i})^{\mathbf{2}}}{\sum(\boldsymbol{yi} - \overline{\boldsymbol{y}})^{\mathbf{2}}}$$

Varie entre [0 ; 1]

R2 exprime le rapport entre la variance expliquée par le modèle sur la variance totale

$R^2 = 0 \Rightarrow$ Le modèle utilisé n'explique pas du tout l'influence de X sur les variations de Y

$R^2 = 1 \Rightarrow$ Le modèle utilisé explique parfaitement l'influence de X sur les variations de Y

$R^2 = 0{,}45 \Rightarrow$ 45% des variations de Y sont expliqués par le modèle utilisé (et donc 65% des variations de Y ont une autre cause

Création d’un modèle

- Fixation de la « graine », pour fixer le processus de randomisation
  - *Reproductibilité*
- Séparation du jeu de données en 2 :
  - *70, 80% pour l’entrainement du modèle*
  - *30, 20% pour le test du modèle*
- Entrainement du modèle
- Evaluation du modèle

Exercice régression linéaire

- Jeu de données : https://raw.githubusercontent.com/john-boyer-phd/TensorFlow-Samples/master/Neural%20Net/bankloanData.csv

- Réaliser une régression avec la somme des dettes : othdebt + creddebt
- Ploter la régression
- Afficher le coefficient R2

- Réaliser une régression linéaire multiple
  - *Split train / test : **sample()***
  - *Entrainer le modèle : **lm()***
  - *Tester le modèle : **predict()***
  - *Calculer la moyenne des écarts : RMSE, MAE (formule à rechercher)*
    - Création de fonctions