SAé 2.03 - Régression sur données réelles

Informations préliminaires

Nom de la SAÉ	**SAé 2.03 - Régression sur données réelles**
Compétence ciblée	Analyser statistiquement les données Niveau 1 : Mettre en œuvre une analyse descriptive
Apprentissages critiques couverts	Comprendre l'intérêt des synthèses numériques et graphiques pour décrire une variable statistique Comprendre l'intérêt des synthèses numériques et graphiques pour mettre en évidence des liaisons entre variables
Description des objectifs de la SAÉ et de la problématique professionnelle associée	L'objectif est de mettre l'étudiant en situation de produire des graphiques et des indicateurs permettant de mettre en évidence la liaison (ou l'absence de liaison) entre deux variables quantitatives. L'étudiant doit prendre en compte le contexte de l'étude (données sociodémographiques, données de consommation, données issues de la santé…), afin de déterminer quels sont les croisements de variables susceptibles d'apporter le plus d'informations. L'étudiant est amené à répondre à une problématique précise du commanditaire de l'étude : comparaison de plusieurs ajustements afin de choisir le plus pertinent, prédiction d'une nouvelle valeur de la variable à expliquer, … Nota : pas de données de type « séries chronologiques » (vu en 2 e année).
Heures formation	10 h dont 10 h de TP
Heures « projet tutoré »	0 h
Liste des ressources mobilisées et combinées	RES 1-05 Statistique descriptive 1 RES 2-03 Programmation statistique RES 2-05 Statistique descriptive 2
Types de livrable ou de production	Rédaction d'un rapport d'étude

Conditions :

- Les travaux seront réalisés en binôme.
- L’évaluation se fera sous forme d’un rapport d’étude (au format PDF).
    - Remarque : Il est fortement conseillé d'utiliser **R** pour à la fois programmer, produire les graphiques et écrire vos analyses et conclusions.

Concepts / notions abordé(e)s : Nuage de point, méthodes des moindres carrés, droite des moindres carrés, ajustement de courbes, part de variance expliquée, études des résidus et prévision.

Objectifs de l'étude

A partir d'un ensemble de données (variables) cliniques relatives à des tumeurs du sein et aux diagnostics associés : **M** (tumeur maligne = tumeur cancéreuse) ou **B** (tumer bénigne = absence de cancer), on souhaite :

- Déterminer, parmi les variables statistiques, les plus discriminantes (c.-à-d. celles dont l'observation permet de distinguer et donc de prédire au mieux l'état malin ou bénin d'une tumeur).
- Rechercher l'existence de corrélations entre les variables statistiques discriminantes retenues.

Support de l'étude

Les données, fournies sous la forme d'un fichier au format CSV, comportent :

- L'identifiant (numérique) de la patiente
- Le diagnostic (**M**: Malin, **B**: Bénin)

Dix variables quantitatives relatives aux noyaux des cellules observées (à partir d'images), à savoir :

1. Le rayon (radius) - Moyenne de distances entre le centre et des points en périphérie;
2. La texture (texture) - Etablie à partir d'une échelle de gris;
3. Le périmètre (perimeter)
4. L'étendue (area)
5. La finesse (smoothness) - Variation locale de longueurs des rayons
6. La compacité (compactness)
7. La concavité (concavity)
8. Les points concaves (concave points) (Nombre de portions concave dans le contour)
9. La symétrie (symmetry)
10. La dimension fractale (fractal dimension)

Les modalités de ces variables quantitatives sont des nombres réels (décimaux)

Pour chaque caractère, on donne :

- La valeur moyenne (mean)
- L'écart-type (standard error)
- Le "pire cas" (mean or largest) - Moyenne des trois plus grandeurs valeurs calculées pour chaque image.

Remarque importante : : Dans cette étude, on ne s'interessera qu'aux valeurs moyennes des caractères.

Travail demandé

Préambule : Tous les traitements de données et tous les graphiques doivent être établis à partir d'une programmation en **R** (ou en Python pour faire plaisir à Monsieur Hébert).

Classification

1. Représenter graphiquement la proportion de tumeurs malignes et de tumeurs bénignes.
2. Tracer puis analyser l'histogramme du rayon (radius_mean)
3. Tracer, sur le même graphique, l'histogramme du rayon (radius_mean) en utilisant deux couleurs différentes pour différencier les tumeurs malignes (M) et des bénignes (B), .
4. Effectuer la même opération que précédemment pour les dix variables statistiques, en positionnant les histogrammes sur deux colonnes et cinq lignes.
5. A partir de la figure précédente et d'une analyse argumentée, en déduire quelles variables pourraient, a priori, être discriminantes pour déterminer l'état de la tumeur.
6. Tracer le nuage de points relatif à la texture en fonction du rayon, en utilisant deux couleurs différentes pour distinguer les tumeurs malignes (M) et bénignes (B). Analyser le nuage de point en termes de corrélation puis par rapport à l'objectif de classification des cas M ou B.
7. Exploiter l'approche graphique précédente pour confirmer ou infirmer le caractère discriminant des variables isolées à la question 5.
8. Si possible, en considérant certaines variables statistiques pertinentes, proposer et mettre en oeuvre une méthode quantitative permettant d'opérer la classification des états M ou B.

Mise en évidence de liaisons entre les variables

Remarque préliminaire : Dans cette partie, on restreindra l'étude aux variables statistiques suivantes :

- Le rayon (radius)
- Le périmètre (perimeter)
- L'étendue (area)
- La compacité (compactness)
- La concavité (concavity)

1. Pour chaque combinaison possible dont on calculera en premier lieu le nombre total, rechercher l'existence de corrélations entre les variables statistiques de la liste précédente.
2. En cas d'une probable corrélation, établir un ajustement linéaire optimal au sens des moindres carrés.

Remarque : Les modèles de régression pourront être fondés sur des transformations logarithmiques des variables.