UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 1 Ingénierie Statistique

Travaux pratiques de Régression

Ex 1. *Lien entre 2 variables quantitatives*
Galilée a réalisé diverses expériences afin de comprendre la trajectoire des boulets de canon. Dans l'une d'entre elles, une balle est lâchée sur un plan incliné et quitte le plan incliné à une certaine hauteur du sol, on mesure ensuite la distance parcourue par la balle. Le tableau suivant résume les données mesurées.

Hauteur	1000	800	600	450	300	200	100
Distance	573	534	495	451	395	337	253

1) Effectuer une représentation graphique de ces données. Un lien apparait-il entre les deux variables ? Semble-t-il linéaire ?

2) Calculer le coefficient de corrélation linéaire entre les deux variables. Est-il significativement non nul ?

3) Même question pour les coefficient de corrélation de Spearman.

4) Quel coefficient quantifie le plus fidèlement le lien observé à la première question ?

Ex 2. *Lien entre 2 variables quantitatives*
Le fichier "lifeexp-TV.dat" contient l'espérance de vie dans différents pays ainsi que le nombre moyen de TV par habitants.

1) Effectuer une représentation graphique des données. Un lien apparait-il entre les deux variables ? Semble-t-il linéaire ?

2) Calculer le coefficient de corrélation linéaire entre les deux variables. Est-il significativement non nul ? Même question avec le coefficient de corrélation de Spearman.

3) Suggérer une transformation des variables rendant le lien "plus" linéaire. Vérifier votre démarche à l'aide d'une représentation graphique et du calcul de la corrélation linéaire.

4) Commenter l'affirmation suivante : "Pour augmenter l'espérance de vie des habitants, il suffit de leur fournir plus de TV".

Ex 3. *Lien entre 2 variables quantitatives*
Le fichier "motordata.txt" contient des enregistrements de l'accélération (en g) de la tête d'un motard après un choc en fonction du temps (en millisecondes).

1) Effectuer une représentation graphique de ces données. Un lien vous semble-t-il apparaître entre l'accélération et le temps écoulé après le choc ?

2) Calculer les coefficients de corrélation de Pearson et de Spearman, et tester leur significativité. Commenter.

Ex 4. *Lien entre 2 variables qualitatives*
La répartition des jeunes âgés de 16 à 25 ans sans diplômes et résidant en Bretagne en 1999 selon leur sexe et leur activité est la suivante (d'après le recensement INSEE) :

Type d'activité	Hommes	Femmes	Total
Actifs ayant un emploi	6 639	2 446	9 085
Chômeurs	2 544	1 850	4 394
Inactifs	1 201	1 872	3 073

1) Saisir ces données sous R et en proposer une représentation graphique synthétique.

2) D'après ces données, peut-on affirmer qu'il y a une dépendance significative entre le type d'activité et le sexe en Bretagne ?

Ex 5. *Lien entre 2 variables qualitatives*
Un site internet reçoit 113 457 visiteurs durant un mois. On désigne par X le navigateur internet utilisé et Y le système d'exploitation utilisé.

X/Y	Windows	Mac	Linux
Chrome	14103	1186	427
Firefox	30853	4392	3234
Internet Explorer	47389	23	0
Safari	668	6416	0
Autres	2974	40	1752

1) Saisir ces données sous R et en proposer une représentation graphique synthétique.

2) D'après ces données, peut-on affirmer qu'il y a une dépendance significative entre le navigateur et le système d'exploitation utilisés ?

Ex 6. *Lien variable quantitative/variable qualitative*
Le fichier "NO2_trafic.csv" contient différentes mesures de dioxyde d'azote (NO2) effectuées au sein de véhicules circulant dans la métropole parisienne. La variable "type" donne le type de routes principalement empruntées ("A" : Autoroute, "P" : Périurbain, "T" : Tunnel, "U" : Urbain, "V" : Voie rapide urbaine) et la variable "fluidite" donne les conditions de trafic (de "A" : fluide, à "D" : congestionné).

1) Proposer une représentation graphique de la variable "NO2" en fonction du type de routes empruntées.

2) Donner la moyenne du NO2 selon le type de routes empruntées.

3) Reprendre les deux questions précédentes pour étudier le lien entre la variable "NO2" et la variable "fluidite".

4) Construire une nouvelle variable de type facteur contenant uniquement la modalité "D" (congestionné) de la variable "fluidite", toutes les autres étant regroupées sous le label "Autre".

5) Y a-t-il une différence significative de la concentration moyenne de NO2 lorsque le trafic est congestionné par rapport aux autres conditions de circulation ? On pourra utiliser un test d'égalité des moyennes à alternative bilatéral et/ou unilatéral.

Ex 7. On veut expliquer la hauteur des eucalyptus en fonction de leur circonférence à partir d'une régression linéaire simple. On dispose des mesures des hauteurs (ht) et circonférences (circ) de 1737 eucalyptus, qui se trouvent dans le fichier "eucalyptus.txt".

1. Extraire et représenter les données dans le plan.
2. Effectuer la régression $y = \beta_0 + \beta_1 x + \epsilon$ où y représente la hauteur et x la circonférence. Commenter les résultats.
3. Calculer un intervalle de confiance à 95% pour β_0 et β_1, en supposant la normalité des données.
4. Si le bruit ϵ ne suit pas une loi normale, les intervalles de confiance précédents restent-ils valables ?
5. Tracer l'estimateur de la droite de régression et un intervalle de confiance à 95% de celle-ci. Que déduisez-vous de la qualité de l'estimation ?
6. On veut à présent prédire la hauteur d'une nouvelle série d'eucalyptus de circonférences 50, 100, 150 et 200. Donner les estimateurs de la taille de chacun d'entre eux et les intervalles de prédiction à 95% associés, en supposant la normalité des données.
7. Si le bruit ϵ ne suit pas une loi normale, les intervalles de prédiction précédents restent-ils valables ?

Ex 8. *Convergence des estimateurs*

I) Lors d'une expérience chimique, on observe la teneur d'un certain produit à différents instants réguliers allant de 1 à n. Le résultat à l'instant i est noté y_i. On suppose le lien temporel suivant : $y_i = \beta_0 + \beta_1 i + \epsilon_i$, $i = 1, \ldots, n$, où les variables ϵ_i représentent les erreurs de mesures. On suppose que $\beta_0 = \beta_1 = 1$ et que les ϵ_i sont i.i.d suivant une loi $\mathcal{N}(0, 20^2)$.

1) Simuler 1000 valeurs de y_i pour i variant de 1 à 1000. Observer le nuage de points entre y et les instants de mesures. Pour k allant de 10 à 1000, effectuer la régression des k premières valeurs de y par rapport aux k premiers instants i. Observer graphiquement l'évolution de l'estimation de β_1 en fonction de k.

2) Répéter la simulation précédente 20 fois et superposer sur un même graphique l'évolution des 20 estimateurs de la pente de la régression. Commenter.

II) On suppose à présent que le lien temporel est $y_i = \beta_0 + \beta_1 \frac{1}{i} + \epsilon_i$ où $\beta_0 = \beta_1 = 1$ et les ϵ_i sont i.i.d suivant une loi $\mathcal{N}(0, 0.1^2)$. Effectuer le même type de simulations que dans la première partie et observer le comportement asymptotique de l'estimateur de β_1.

Ex 9. *Consommation de glaces*

On étudie la consommation de glaces aux Etats-Unis sur une période de 30 semaines du 18 Mars 1950 au 11 Juillet 1953. Les variables sont la période (de la semaine 1 à la semaine 30), et en moyenne sur chaque période : la consommation de glaces par personne ("Consumption", en 1/2 litre), le prix des glaces ("Price", en dollars), le salaire hebdomadaire moyen par ménage ("Income", en dollars), et la température ("Temp", en degré Fahrenheit). Les données sont disponibles dans le fichier "icecream-R.dat".

1) Extraire les données et représenter la consommation en fonction des différentes variables.

2) On propose de régresser linéairement la consommation sur les trois variables "Price", "Income" et "Temp", en supposant de plus qu'une constante est présente dans le modèle. On note la constante β_1 et les trois coefficients associés aux variables précédentes respectivement β_2, β_3 et β_4. Réaliser la phase d'estimation de cette régression et commenter le signe des coefficients estimés.

3) Tester la significativité globale du modèle proposé, i.e. $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$, à l'aide du test de Fisher global.

4) Tester la significativité de la variable "Price" dans ce modèle au seuil de 5%. Tester de même la significativité de "Income", puis de "Temp".

5) Comparer le modèle complet précédent et le modèle sans la variable "Price" à l'aide d'un test de Fisher :
1. En basant le calcul sur la somme des carrés résiduelle de chaque modèle ;
2. En basant le calcul sur le coefficient de détermination de chaque modèle ;
3. En utilisant la fonction `linearHypothesis` de la librairie `car`.
4. En utilisant la fonction `anova`.
Quel est la différence entre ce test et le test de Student de significativité de la variable "Price" ?

6) Comparer le modèle complet et le modèle sans la variable "Price" et sans la constante à l'aide d'un test de Fisher. Procéder selon les 4 manières décrites ci-dessus. Commenter.

7) On désire à présent prédire la consommation de glaces pour les données suivantes : $Price = 0.3$, $Income = 85$ et $Temp = 65$. Proposer la prévision qui vous semble la meilleure au vu des modèles étudiés précédemment. Donner un intervalle de prédiction au niveau 95% autour de cette prévision.

8) Sous quelle hypothèse l'intervalle de prédiction précédent est-il valable ? Vérifier-la en observant le QQ-plot des résidus de la régression et en effectuant un test statistique.

9) Vérifier les autres hypothèses en rappelant la définition et en calculant les VIF ("Variance Inflation Factor") de chaque variable explicative et en effectuant une analyse graphique des résidus.

10) Observer le nuage de points en 3 dimensions des variables "Consumption", "Income" et "Temp", et l'ajustement par le modèle linéaire, à l'aide de la fonction `scatter3d` de la librairie `car`.

Ex 10. *Emissions de Gaz à effet de serre*

Le jeu de données "EmissionsGES.txt" contient les émissions de Gaz à effet de serre (GES) pour l'année 2003 dans 16 pays industrialisés (en millions de tonnes équivalent CO2). Il contient également les objectifs d'émissions pour 2010 issus des accords de Kyoto (pour les pays ayant ratifié le protocole).

1) Ajuster un modèle régression linéaire expliquant les objectifs d'émissions de GES en 2010 en fonction des émissions mesurées en 2003. Analyser la qualité du modèle. D'après l'estimation, quel est le pourcentage moyen envisagé de réduction des émissions de GES ?

2) Ajuster un modèle de régression linéaire expliquant les émissions de GES en 2003 en fonction de la population des pays. Analyser la qualité du modèle, notamment l'influence des individus (c'est à dire les pays) sur l'estimation à l'aide de leur distance de Cook.

3) Effectuer le même ajustement que ci-dessus sans prendre en compte les USA. Analyser la qualité du modèle.

Ex 11. *Modélisation de la concentration maximale journalière en ozone*

Le jeu de données "ozone.txt" contient la concentration maximale d'ozone (maxO3) mesurée chaque jour de l'été 2001 à Rennes. Il contient également les températures, la nébulosité et la vitesse du vent mesurés à 9h, 12h et 15h (respectivement T9, T12, T15, Ne9, Ne12, Ne15 et Vx9, Vx12, Vx15), ainsi que la direction principale du vent et la présence ou non de pluie. On désire expliquer au mieux la concentration d'ozone à l'aide des variables disponibles dans le jeu de données.

1) Analyser le nuage de points et la corrélation linéaire entre maxO3 et chacune des variables quantitatives disponibles (c'est à dire T9, T12, T15, Ne9, Ne12, Ne15, Vx9, Vx12 et Vx15). Est-il raisonnable de supposer qu'il existe un lien linéaire entre maxO3 et ces variables ?

2) Ajuster le modèle de régression linéaire expliquant maxO3 en fonction de toutes les variables quantitatives précédentes. Tester la significativité de chacune des variables explicatives dans ce modèle. Le résultat est-il en accord avec les observations de la question précédente ?

3) Calculer les VIF (Variance Inflation Factor) pour chacune des variables explicatives du modèle précédent. En quoi ces valeurs expliquent les résultats des tests de Student effectués ci-dessus ?

4) On décide d'enlever des variables au modèle précédent. Quelles variables semble-t-il naturel d'enlever au vu de la question précédente ? Ajuster le ou les nouveaux modèles proposés et répéter les analyses effectuées dans les deux questions précédentes.

5) Sélectionner le meilleur modèle possible ayant toutes ses variables significatives et aucun problème de multicolinéarité. Le choix du modèle pourra reposer sur un critère de sélection de type BIC.

6) Mettre en oeuvre une sélection automatique du meilleur sous-modèle possible du "gros" modèle ajusté dans la question 2, selon le critère BIC. On pourra utiliser la fonction `regsubsets` dans la libraire `leaps` (puis `plot.regsubsets`) ou `step`. Comparer le modèle retenu avec le modèle choisi à la question précédente.

7) Appliquer la sélection automatique précédente en vous basant sur d'autres critères que BIC. Les modèles retenus sont-ils les mêmes ? Si non, lequel semble préférable ?

8) Analyser résidus du modèle sélectionné à la question précédente par des représentations graphiques et en effectuant des tests d'homoscédasticité et de non-corrélation des résidus. Toutes les hypothèses d'un modèle linéaire semblent-elles vérifiées ?

9) Afin de résoudre le problème d'auto-corrélation des résidus, on propose d'ajouter la maximum d'ozone de la veille dans le modèle. Créer cette variable, que l'on nommera maxO3v et ajouter-la au jeu de données. Observe-t-on un lien linéaire entre maxO3 et maxO3v ?

10) Ajuster le modèle de régression contenant maxO3v comme variable explicative supplémentaire. Analyser les résultats de l'ajustement : les hypothèses d'un modèle linéaire sont-elles vérifiées ?

11) Comparer ce dernier modèle au modèle sans maxO3v à l'aide d'un test de Fisher et en comparant les différents critères de sélection (AIC, BIC, Cp de Mallows, R^2 ajusté).

Ex 12. *ANOVA à 1 facteur : effet du dosage d'un médicament*
Le fichier "chemical.txt" contient la concentration dans le sang (en ng/ml) d'un certain produit chimique chez 40 patients selon qu'ils ont absorbé un médicament dosé à 25, 50, 100 ou 200 mg de substance active (l'almitrine bismesylate). Les patients sont ainsi séparés en 4 groupes de 10 selon le dosage reçu.

1) Représenter les données à l'aide de boîtes à moustaches. Réordonner si nécessaire les niveaux du facteur "dose" pour avoir une représentation par ordre croissant du dosage. Les hypothèses nécessaires à une analyse de variance de l'effet dosage semblent-elles vérifiées ?

2) Effectuer un test d'égalité de variance de Bartlett (`bartlett.test`) et de Levene (`leveneTest` dans la librairie `car`) pour confirmer le problème vu précédemment.

3) On propose de s'intéresser à une transformation logarithmique de la variable "concent". Représenter les boîtes à moustaches des données transformées. Effectuer des tests d'égalité des variances pour vérifier la stabilité des variances après transformation.

4) Réaliser l'analyse de la variance et tester l'effet du dosage sur la concentration en produit chimique.

5) Analyser graphiquement les résidus du modèle pour valider la démarche.

6) On veut à présent comparer plus précisemment les effets des dosages entre eux. Combien de paires de dosages cela fait-il à comparer ? Effectuer pour chaque paire un test de Student de comparaison des moyennes au niveau $\alpha = 0.05$ (fonction `t.test`).

7) La sortie de la fonction `t.test` fournit également un intervalle de confiance pour la différence des espérances entre les 2 groupes testés. Peut-on affirmer qu'avec une probabilité de 0.95 toutes les différences d'espérances appartiennent aux intervalles de confiance obtenus ci-dessus par la fonction `t.test` au niveau $\alpha = 0.05$?

8) Proposer de même des intervalles de confiance au niveau simultané 0.95 en utilisant la méthode de Tukey (fonction `TukeyHSD` ; si le modèle se nomme `reg` : `TukeyHSD(aov(reg))`) et représenter les (`plot` du résultat).

Ex 13. *ANOVA à 2 facteurs : alimentation des rats*
On veut étudier l'évolution du poids des rats selon 4 régimes alimentaires différents : combinaison de deux types de protéines différentes (Boeuf et Céréales) et de deux quantités différentes (élevée ou basse). Chaque traitement est effectué sur 10 rats choisis au hasard de telle sorte que notre échantillon est constitué de 40 rats. Les données se trouvent dans le fichier "poids-rats.txt".

1) Le plan est-il équilibré ?

2) Représenter les données à l'aide de 4 boîtes à moustaches croisant le poids et les deux facteurs. Les conditions semblent-elles remplies pour effectuer une ANOVA ? Effectuer un test.

3) Analyser l'effet de l'interaction des deux facteurs sur le poids à l'aide d'un graphique (utiliser la fonction `interaction.plot`) puis effectuer le graphique équivalent en échangeant le rôle joué par les facteurs. Commenter.

4) Réaliser l'analyse de la variance complet à deux facteurs (incluant les effets marginaux et l'interaction entre les facteurs) et tester la présence d'une interaction entre les facteurs. Comparer ce résultat avec les représentations précédentes.

5) On considère à présent le modèle additif sans interaction. Estimer ce modèle. Sous quelles contraintes sur les paramètres cette estimation est-elle effectuée ?

6) Tester l'effet de chacun des facteurs.

7) Effectuer les tests multiples de Tukey d'égalité des moyennes entre chaque modalité croisée.

Ex 14. *ANOVA à 2 facteurs : effets sur la perte de poids*
Le fichier "Diet.txt" résume la perte de poids de 72 personnes ayant suivi pendant 6 mois un régime alimentaire particulier (parmi 3 possibles) et un programme d'activité physique spécifique (parmi 4 possibles). Analyser l'effet du régime et/ou de l'activité physique sur la perte de poids.

Ex 15. *ANCOVA : Retour à la modélisation de l'ozone*
On considère de nouveau les données "ozone.txt" étudiées dans l'exercice 11 de TP.

1) On reprend le modèle sélectionné dans l'exercice 11, soit "maxO3" en fonction de "T12", "Ne9", "Vx9" et "maxO3v" où "maxO3v" représente la concentration maximale en ozone de la veille (créer cette variable si besoin). Ajuster ce modèle sur les données.

2) Représenter graphiquement "maxO3" en fonction de la présence de pluie. Un lien semble-t-il présent ?

3) Ajouter au modèle de la première question la variable "pluie" de manière la plus générale possible (i.e. en incluant une interaction avec chaque variable en plus d'un effet sur la constante). Tester la significativité de ces ajouts en effectuant un test de Fisher de modèles emboités entre ce modèle et le modèle initial.

4) Tester de même le modèle plus simple dans lequel seul un effet additif de la variable "pluie" est intégré, et non ses interactions avec les autres variables. Le résultat est-il en désaccord avec l'analyse graphique précédente ? Comment expliquer le résultat ?

5) De même : représenter graphiquement "maxO3" en fonction de la direction du vent et étudier la pertinence d'inclure un effet vent dans le modèle initial.

Ex 16. *Régression logistique : maladie coronarienne*
Le jeu de donnée "chdage.txt" contient 100 patients de 20 à 69 ans (variable "age"), présentant pour certains une maladie coronarienne (variable "chd").

1) Représenter l'âge des patients en fonction de la présence ou non de la maladie. Y-a-t-il une différence significative ? Effectuer un test.

2) Recoder la variable "chd" en 0/1, puis représenter-la en fonction de la variable "age".

3) Calculer la proportion de malades dans chaque classe d'âge de la variable "agegr" (on pourra utiliser les fonctions `table` et `prop.table`). Superposer au graphe précédent ces proportions en prenant comme abscisses les centres de chaque classe.

4) Ajuster un modèle de régression logistique expliquant $\pi(x)$, la probabilité d'être atteint d'une maladie coronarienne en fonction de l'âge x (fonction `glm` avec l'option `family="binomial"`).

5) L'âge est-il une variable significative dans le modèle précédent ? Quel test permet de répondre à cette question ?

6) D'après cet ajustement, à quelle valeur peut-on estimer la probabilité d'être atteint d'une maladie coronarienne à 30 ans ? à 60 ans ?

7) Superposer la courbe $\hat{\pi}(x)$ au graphe précédent.

8) Calculer les intervalles de confiance au niveau 95% pour $\pi(x)$ et les superposer au graphe précédent (utiliser la fonction `predict` avec l'option `se=T`).

Ex 17. *Régression logistique : survivants du Titanic*
Le jeu de données "Titanic.txt" renseigne pour chaque passager du Titanic, sa classe (crew, 1ère, 2nde ou 3ème), son âge (enfant ou adulte), son sexe, et s'il a survécu ou non au drame.

1) Modéliser la probabilité de survie en fonction de toutes les variables à disposition, en incluant éventuellement des interactions entre elles. On testera la significativité de chaque facteur par un test de déviance (fonction `anova` avec l'option `test="Chisq"`). On sélectionnera le meilleur modèle par un test de déviance entre modèles (même fonction) et par comparaison des critères BIC (fonction `BIC`).

2) Reprendre la question précédente en effectuant une sélection automatique du meilleur modèle à l'aide de la fonction `step`, basée sur le critère BIC, partant du modèle le plus gros possible (contenant toutes les interactions).

3) Analyser l'estimation du modèle retenu.

Ex 18. *Régression logistique : modes de transport*
Le fichier "Mode_app4.txt" contient, pour différents trajets effectués, le mode de transport choisi (entre "voiture", "covoiturage", "bus" et "train"), ainsi que la durée et le prix de chaque transport.

1) On s'intéresse dans un premier temps au transport par le train. Créer une nouvelle variable binaire codant le mode de transport "train" ou "autre" .

2) Effectuer une régression logistique expliquant la probabilité de prendre le train en fonction des variables à disposition. On sélectionnera le meilleur modèle par une sélection stepwise descendante, ou selon le critère BIC.

3) En utilisant le modèle retenu, estimer la probabilité précédente pour chaque déplacement par validation croisée (par leave-one-out).

4) Tracer la courbe ROC issue de ces prévisions par validation croisée (fonction `performance` de la librairie `ROCR` et la validation croisée se fait "à la main").

5) On décide de prédire un déplacement "train" si la probabilité estimée est supérieure au seuil s. Choisir le seuil s (toujours par validation croisée) qui minimise le nombre total de mauvais classés (option `measure="err"` de `performance`).

6) Le fichier "Mode_valid4.txt" contient les données pour des nouveaux trajets. Utiliser la stratégie retenue pour classer ces trajets selon qu'ils utilisent le transport "train" ou "autre". Dresser la matrice de confusion.

7) On désire à présent prédire les 4 modes de transport possibles. Effectuer une régression logistique multinomiale (fonction `multinom` de la librairie `nnet`) à partir du fichier "Mode_app4.txt" en sélectionnant le meilleur modèle selon le critère BIC (fonction `stepAIC`).

8) Prédire le mode de transport des trajets du fichier "Mode_valid4.txt" en utilisant le modèle sélectionné. Dresser la matrice de confusion.