TP Statistiques et Séries Chronologiques
Université de Lorraine

Régression Linéaire Multiple et Initiation aux tests statistiques

Clément Dell'Aiera

1 Exercice 1

1. Que font les commandes *pnorm*, *qnorm* , *rnorm*, *dnorm* ? Utilisez l'aide de *R*.
2. Tracer la fonction de densité de la loi normale. Faites varier les paramètres et afficher les différentes courbes sur le même graphique.
3. Simuler 2 vecteurs X et Y contenant chacun $N = 100$ variables indépendantes identiquement distribuées suivant une loi normale $\mathcal{N}(0, 1)$.
4. Afficher les points de coordonnées $(X[j], Y[j])$ dans le plan, pour j allant de 1 à 100.
5. Tracer la fonction de répartition empirique des $X[j]$.
6. Soit $\mathcal{E}$ une v.a. de loi exponentielle de paramètre 1, et U une v.a. suivant une loi uniforme sur $[0, 2\pi]$. On pose

$$(X, Y) = (\sqrt{\mathcal{E}} \cos(U), \sqrt{\mathcal{E}} \sin(U)).$$

   Quelle est la loi du couple (X, Y) ? (Vous pouvez le prouvez, ou observer grâce à *R* ce qu'il se passe en simulant ces variables et en les traçant.)

2 Exercice 2

1. Une table est déjà en mémoire dans *R* : la table *stackloss*. Analyser la rapidement.
2. Tracer *stack.loss* en fonction de *Air.Flow*. Qu'en pensez-vous ?
3. Effectuer la régression linéaire de *stack.loss* en fonction des autres variables. Quelles sont celles qui sont significatives ?

3 Exercice 3

Le fichier *ozone.dta* contient les variables suivantes, pour une série de journées (qui sont ici nos individus) :

- l'identifiant de la journée,
- le maximum d'ozone (variable maxO3)
- l'heure à laquelle le maximum d'ozone a été obtenu (heure),
- les températures à 6h, 9h, 12h, 15h, 18h (resp. T6 à T18)
- la nébulosité à 6h, 9h, 12h, 15h, 18h (resp. Ne6 à Ne18)
- la projection du vent sur l'axe est-ouest à 12h (Vx),
- le maximum d'ozone de la veille (maxO3v).

Le but est de modéliser la valeur des pics d'ozone en fonction de grandeurs physiques facilement mesurables (température, heure, nébulosité, vent) afin d'avoir des approximations de la qualité de l'air faciles et rapides à obtenir.

1. Importer la table, et afficher un résumé de ce qu'elle contient.
2. Tracer *maxO3* en fonction de *T12*, puis effectuer un régression linéaire. Ajouter la droite de régression sur le graphique. Soignez la présentation.
3. Afficher les résultats de la régression.

4. Extraire les résidus et tracer leur densité estimée.
5. Effectuer la régression de *maxO3* sur toutes les variables, et supprimer récursivement celles qui ne sont pas significatives, jusqu'à ce qu'elles le soient toutes.

4 Exercice 4

1. Sur le site *data.gouv.fr*, vous pourrez trouver des tables de données publiques en libre accès. Choisissez un thème qui vous intéresse, puis une table en conséquence. Télécharger là.
2. Les tables sont souvent au format *.xls* : vous aurez besoin d'installer un package pour pouvoir les lire. La commande pour ce faire est *install.packages("nom du package")*. Installer le package *gdata*.
3. Analyser votre table.

5 Exercice 5

1. Télécharger et installer le package *ISwR* (Introductory Statistics with R).
2. Utiliser la commande *summary* pour analyser rapidement la table *bp.obese*. L'échantillon provient d'un échantillon de population mexicaine en Californie, et la table décrit 3 variables : le sexe (femme = 1, homme =0), le ratio d'obésité (*obese*) et la pression sanguine systolique en *mm* de mercure (*bp*).
3. Représenter les données dans un graphe, en utilisant des symboles différents pour les hommes et les femmes.
4. Expliquer la pression sanguine en fonction du ration d'obésité, puis du ratio d'obésité et du sexe.
5. Tracer sur un même graphe les courbes correspondant aux régressions dans les 2 modèles. Soignez la présentation (couleurs différentes, légende,...)

6 Exercice 6

1. Télécharger la librairie *MASS*.
2. Analyser rapidement la table *cats* et afficher les variables les unes en fonctions des autres par paires.
3. Effectuer une régression linéaire selon le modèle $Hwt \sim Bwt * Sex$. Cela apporte-t-il quelque-chose par rapport au modèle $Hwt \sim Bwt + Sex$?
4. Visualiser les composantes de votre régression.
5. En extraire les prédiction, les coefficients, les résidus, lés résidus studentisés, et la formule du modèle.
6. Tracer le *qqplot* des résidus studentisés ainsi que la première bissectrice.
7. Tracer le graphe des résidus contre les prédictions.
8. Tracer le graphe des distance de Cook.
9. Observer les attributs que vous donne *summary*

10. Afficher le R^2 ajusté de la régression, le nombre de degrés de liberté résiduels, la matrice de variance-covariance des paramètres estimés.

7 La régression linéaire sous R

Ce numéro rappelle les notions nécessaires à l'interprétation d'une sortie R de la fonction *lm*. Dans toute la suite, *regression* désigne un objet de type *lm* que l'on a appellé grâce à $lm(Y \sim X_1 + ... + X_n, data = ...)$. Si l'on note X_{-i}, le moins signifie que l'on calcule la quantité X sans tenir compte de l'observation i.

Dans le modèle de régression

$$Y = X\beta + \epsilon,$$

la commande *summary(regression)*, où *regression* est un objet de la classe lm, renvoie plusieurs tableaux.
On note $\hat{y}_j = \sum h_{ij}x_j$, i.e.

$$h_{ij} = \frac{1}{n} + \sum \frac{(x_i - \hat{x})(x_j - \hat{x})}{\sum (x_j - \hat{x})^2}.$$

Le premier tableau,*Residuals*, est destiné à donner une idée de la répartition des résidus en affichant les quantiles. Je vous recommande d'afficher tout de même les résidus, et d'observer leur distribution. Mieux, vous pouvez utiliser les résidus studentisés. A priori, bien que tous soient centrés, les résidus n'ont pas même variance (même sous hypothèse d'homoscédasticité !) : $Var[\epsilon_j] = \sigma^2(1 - h_{jj})$. Pour les rendre comparables, on pourrait les réduire, mais si l'on remplace la variance par la variance estimée, celle-ci dépend de l'information contenue dans x_j, ce qui empêche une quantification de l'effet que x_j a sur les coeffcients de la régression. Pour palier à ce problème, on introduit

$$\hat{\sigma}^2_{-j} = \frac{1}{n-3}[(n-2)\hat{\sigma}^2 - \frac{\epsilon_j}{1 - h_{jj}}]$$

qui n'est rien d'autre que la variance estimé sur le modèle où l'on a supprimé l'observation x_j. Les résidus studentisés sont définis par $T_j = \frac{\epsilon_j}{\sigma_{-j}(1-h_{jj})} \sim T(n-3)$ et suivent une loi de Student à $n-3$ degré de liberté sous des hypothèses raisonnables. Pour détecter une anomalie dans les données, on peut vérifier que les résidus studentisés se répartissent de manière uniforme sur l'intervalle $[-2; 2]$ (sous hypothèse d'homoscédasticité). Repérer des formes suspectes est un moyen facile pour repérer les valeurs aberrantes. On peut par exemple taper :
qqnorm(studres(regression)) ; qqline(studres(regression)).
Une autre méthode pour détecter les valeurs aberrantes : utiliser la distance de Cook. Elle est définie par

$$D_j = \frac{\sum_j (\hat{y}_{-i,j} - \hat{y}_j)^2}{2\hat{\sigma}^2},$$

et mesure l'influence d'une observation sur l'ensemble des prévisions (qu'on veut petite !). Encore une règle du pouce : si $D_j > 1$, on enlève l'observation j. Vous pouvez le faire automatiquement en tapant *plot(regression,which = 4)*.
Le deuxième est nommé *Coefficients* :

	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
β_j	$\hat{\beta}_j$	$\hat{\sigma}_j$	$\hat{t}_j = \frac{\hat{\beta}_j}{\hat{\sigma}_j}$	p-value

Les trois premières colonnes s'expliquent elles-mêmes, mais à quoi servent les 2 dernières ? A effectuer un test de significativité. Plus précisément, on sait que $\hat{t}_j$ suit une loi de Student à $N-k$ degrés de liberté sous l'hypothèse $H_0|\beta_j = 0$ contre $H_1|\beta_j \neq 0$. La quatrième colonne donne donc la valeur de cette statistique, et la dernière sa p-value, définie comme la valeur seuil de confiance α qui fait basculer le test. (Rappelez vous que, mécaniquement, si α diminue assez, on finit par accepter H_0.) Une règle appliquée par les statisticiens est la suivante :

$p < 0.01$	suspicion très forte contre H_0
$0.01 - 0.05$	suspicion forte contre H_0
$0.05 - 0.1$	suspicion faible contre H_0
> 0.1	peu ou pas de suspicion contre H_0

Donc, si $Pr(> |t|)$ est petit, on rejette l'hypothèse $\beta_j = 0$, ce qui signifie que le coefficient est significatif. R ajoute même des petites étoiles à côté des coefficients les plus significatifs.

Reste encore à observer plusieurs indicateurs. L'erreur *Residual standard error* est calculée comme un estimateur de σ sous l'hypothèse de matrice variance-covariance égale à $\sigma^2 I_n$. Il nous reste encore le R^2 et le R^2 ajusté. Rappelons que le coefficient R^2 peut s'interpréter comme le cosinus de l'angle entre le vecteur des observations Y_j et son projeté pour la norme $\mathcal{L}^2$ sur l'espace linéaire engendré par les observations X_j, soit

$$R^2 = \frac{\sum(\hat{y}_j - \overline{y})^2}{\sum(y_j - \overline{y})^2}.$$

Cet indicateur à des valeurs comprises entre 0 et 1, la proximité avec 1 indiquant une bonne adéquation du modèle aux données. Toutefois, son interprétation est sujette à caution : sa valeur augmente mécaniquement avec l'ajout de variables explicatives. En particulier, pour comparer la qualité de deux modèles au nombre de variables explicatives distinct, on lui préférera le R^2 ajusté, qui prend en compte ce nombre noté k :

$$RR^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

8 Théorie des tests

8.1 Exemples

1. Une entrerise vend des biens dont elle assure que la durée de vie dépasse 10000 heures. Vous êtes engagés pour vérifier la qualité d'iceux. Vous disposez d'un échantillon de 30 de ces biens. La durée de vie moyenne calculée sur l'échantillon vaut 9900 heures, on suppose que l'écart-type est connu et vaut 120 heures. Pouvez-vous rejeter leur assertion à à un niveau de confiance de 0.05%.
   Répondez à la même question si l'écart-type n'est plus connu, et que l'écart-type obtenu sur l'échantillon vaut 125 heures.
   Calculez la puissance du test, c'est-à-dire la probabilité de l'erreur de seconde espèce.
2. Une firme agroalimentaire assure qu'un cookie qu'elle produit ne contient pas plus de 2 grammes d'un certain composé (graisse, colorant,...). Vous achetez un paquet, contentant 35 cookiees, et mesurez une teneur moyenne de 2.1 grammes. En supposant que l'écart-type de l'échantillon est de 0.25, pouvez-vous incriminez la firme à un niveau de confiance de 0.05%. Même question si l'on ne connaît que l'écart-type empirique de 0.3. Calculez la puissance du test.
3. Lors des dernières élections, les médias affirment qu'au moins 60% des citoyens ont voté. Vous interrogez 148 citoyens de façon à obtenir un échantillon représentatif de la population (vous êtes statisticien après tout). Vous obtenez que 85 des personnes interrogées ont voté. A 0.05%, votre test concorde-t-il avec l'affirmation des médias ?

8.2 Analyse of Variance ou procédure ANOVA

1. Une institution de santé publique veut comparer l'effet de trois traitement contre la grippe. Pour cela, 18 hopitaux sont choisis de façon aléatoire, répartis par groupes de 6, chacun appliquant un et un seul des trois traitement. Voici le nombre de personnes guéries au bout d'une semaine de traitement :

Trait. 1	Trait. 2	Trait. 3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39

   (a) Une méthode pour rentrer les données dans R : les taper dans un fichier *.txt* puis utiliser *read.table* pour créer un objet de la classe *data frame.* (Faîtes-le)
   (b) Utiliser un test ANOVA pour répondre à la problématique de l'institution.

(c) Un pays frontalier, lui aussi touché par l'épidémie, décide de répliquer l'expérience avec le même nombre d'hôpitaux et les mêmes traitements. Chaque hôpital est selectionné de façon aléatoire, et dois appliquer les trois traitements pendant trois semaines, chaque traitement pendant une semaine, l'ordre des traitements étant lui aussi aléatoire. Voici le résultat :

Trait. 1	Trait. 2	Trait. 3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39

9 Régression sur variables qualitatives

Le but de cet exercice est d'expliquer la concentration en ozone *O3* en fonction de la température *T12* et de la direction du vent *vent* dans la table *ozone.txt*.

1. Télécharger la table, et effectuer des régressions selon les différents modèles.
2. Tester l'égalité des pentes.
3. Tester l'égalité des ordonnées à l'origine.
4. Analyser les résidus.

9.1 ANOVA à 1 facteur

Nous souhaitons modéliser la concentration en ozone en fonction de la direction du vent.

1. Tracer une boîte à moustaches de la variable $O3$ par rapport aux quatres modalités de la variable *vent*. Le vent semble-t-il avoir une influence sur la concentration en ozone ?
2. On se place dans un modèle d'analyse de la variance à un facteur

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

   (a) Effectuer la regression linéaire de *O3* sur *vent* sous la contrainte $\mu = 0$.
   (b) Effectuer la regression linéaire de *O3* sur *vent* sous la contrainte $\alpha_1 = 0$.
   (c) Effectuer la regression linéaire de *O3* sur *vent* sous la contrainte $\sum n_i \alpha_i = 0$.
   (d) Effectuer la regression linéaire de *O3* sur *vent* sous la contrainte $\sum n_i \alpha_i = 0$.
3. Analyser les résidus afin de constater que l'hypothèse d'homoscédasticité est vérifiée. Pour cela, tracer un boxplot des résidus en fonction de *vent*, les résidus en fonction de $\hat{O}3$, leurs quantiles théoriques ainsi que la distribution des résidus par modalité de *vent*.

9.2 ANOVA à 2 facteurs

Nous voulons maintenant modéliser la concentration en ozone par le vent et la nébulosité, variable à 2 modalités : SOLEIL et NUAGEUX.

1. Procéder à un examen graphique qui puisse déterminer si l'interaction des facteurs influe sur la variable à expliquer. (voir ce qu'est un *profil*)
2. On suppose la gaussianité des résidus.
   (a) Tester le modèle avec interaction : **mod1**.
   (b) Tester le modèle sans interaction : **mod2**.
   (c) Tester le modèle sans effet du facteur *nebulosité* : **mod3**.
3. Grâce à la commande ANOVA de R, effectuer des analyses de la variance entre les modèles **mod1**, **mod2** et **mod3**.
4. Répondez à la problématique.

10 Follicule de rose

Cet exercice est fortement inspiré d'un exercice du livre *Initiation à la statistique avec R* de F. Bertrand et M. Maumy-Bertrand aux édition Dunod. Vous renderez une copie sur laquelle chaque question sera détaillée, et vous enverrez votre fichier *.R* à mon adresse mail ainsi que tous vos graphiques. Votre fichier *.R* devra être nommé *NomPrénom.R*. Lorsque l'on demande de donner des conclusions, cela signifie que l'on s'attend à une **phrase rédigée** sur la copie, ainsi que du code **commenté** dans le fichier *.R*. De plus, les graphiques et autres tableaux ne peuvent qu'être bénéfiques à la compréhension, et donc à la notation, de votre travail.

Cet énoncé s'intéresse à quantifier l'effet de la taille des follicules de roses sur leurs masse, ainsi que l'effet possible de l'espèce. Dans la première partie, on ne s'intéresse qu'à la relation entre taille et masse, la seconde essaie d'incorporer l'espèce (qui est une variable qualitative).

1. Télécharger le package *BioStatR* et importer la librairie ainsi que la table *Mesures*. Décrire rapidement la table *Mesures*. (Nombre de variables, principales caractéristiques des variables,etc.)
2. Tracer le nuage de points de la taille en fonction de la masse des follicules de roses.
3. Pouvez-vous soupçonner une relation linéaire entre ces deux variables ?
4. Effectuer la régression linéaire de la taille sur la masse. Tracer la droite de régression sur le graphique. Soignez la présentation (titre, couleurs,etc).
5. Donner un intervalle de confiance pour les coefficients du modèle que vous obtenez. Comment avez-vous trouvé ces informations ? En bonus, vous pouvez tracer la zone de confiance pour la droite de régression sur votre graphique.
6. Donnez une prédiction plausible de la taille d'un follicule si sa masse est de 42 grammes.
7. Comment appellez-vous l'écart entre la valeur observée d'une observation et celle prédit par le modèle ?
8. La droite des moindres carrés passe-t-elle par le point moyen $(\overline{x}_n, \overline{y}_n)$? Si oui, démontrer-le.
9. Donner le coefficient de détermination R^2. A quoi sert ce coefficient ? Quelle est la différence entre le R^2 et le R^2 ajusté ?
10. Donner la part de variance expliquée et totale.
11. Donnez la valeur de l'estimation de σ^2.
12. Effectuer les deux tests suivants au risque 0.05% :
    — $H_0 \mid \beta_0 = 0$ contre $H_1 \mid \beta_0 \neq 0$.
    — $H_0 \mid \beta_1 = 0$ contre $H_1 \mid \beta_1 \neq 0$.

Cette partie vise à quantifier l'effet de l'espèce sur la taille du follicule.

1. Effectuer une régression croisée qui tient compte de la variable *espece*. Donner une synthèse rapide du résultat.
2. Comment pourriez-vous tester si l'espèce à un effet sur la taille ?
3. Proposer une méthode, le mettre en oeuvre et donnez vos conclusions.

11 Problème du voyageur de commerce et algorithme du recuit simulé

D'après le livre de Michel Benaïm et Nicole El Karoui, *Promenade aléatoire, Chaîne de Markov et simulations, martingales et stratégies*, exemple 3.1.8.

La méthode du recuit simulé est un algorithme d'optimisation proche de celui de Metropolis, et consiste à se promener aléatoirement sur l'espace (fini) des états d'un système selon une loi construite de façon à ce que la promenade converge vers un état qui minimise une certaine fonctionnelle.

On se donne une fonction $h :]0;\infty[\to]0;1]$ telle que

$$h(x) = xh(\frac{1}{x}),$$

par exemple $min(1,x)$ ou bien $\frac{x}{1+x}$. La fonction $V : E \to \mathbb{R}_+$ est la fonction "coût" à minimiser.

La terminologie "recuit simulé" vient d'une technique métallurgique consistant à faire fondre de façon répétée le métal puis à le faire lentement refroidir pour en améliorer les propriétés. En effet, on va se donner un schéma de décroissance d'une quantité analogue à la température, que nous noterons T_n. Ce schéma est crucial pour que l'algorithme converge vers un minimum global de la fonction V et ne reste pas piégé dans un de ses minima locaux. Vous pourrez choisir l'un des schémas suivants :

— décroissance logarithmique :

$$T_n = \frac{C}{\log(n)}$$

— recuit par palier :

$$T_n = \frac{1}{k} \text{ pour } e^{(k-1)C} \leq n < e^{kC}$$

Voici l'algorithme :

> Initialiser X_0. Choisir le nombre de pas de la marche aléatoire m. Pour n allant de 1 à $m-1$, répéter :
>
> 1. Choisir un voisin y de X_n aléatoirement.
> 2. Tirer $U \sim \mathcal{U}_{[0,1]}$.
> 3. Si $U < h(\exp(\frac{1}{T_n})(V(X_n) - V(y))\frac{N(y)}{N(X_n)})$, accepter $X_{n+1} = y$, sinon refuser i.e. $X_{n+1} = X_n$.

Le problème auquel nous allons appliquer cet algorithme est celui d'un commerçant devant visiter un ensemble fini $E = X_1, ..., X_N$ de villes, une et une seule fois. Pour minimiser son temps et son argent, il souhait trouver le chemin l qui minimise la distance, soit, avec nos notations :

$$V(l) = \sum_{j=1}^{N-1} d(X_{l(j)}, X_{l(j+1)}),$$

où l'on voit un chemin comme une permutation de l'ensemble E. Nous travaillerons avec des villes disposées aléatoirement dans le carré $[0;1] \times [0;1]$.

1. Créer une fonction qui calcule le coût d'un chemin donné l.
2. Créer une fonction qui affiche un chemin donné l.
3. Choisir une loi de transition sur les chemins, et l'implémenter.
4. Implémenter l'algorithme du recuit simulé sur ce problème, à l'aide d'une fonction si possible. Afficher le chemin obtenu, ainsi que l'évolution de la longueur du chemin en fonction du nombre d'itérations . Qu'en pensez vous ? De combien d'itérations avez-vous besoin pour obtenir un chemin plausible ?