ENSC 2A - "Modélisation statistique"

TP 2 : Régression linéaire multiple avec le logiciel R

1 Introduction très brève

L'objectif principal de la régression linéaire multiple est d'expliquer une variable quantitative y à partir de p variables explicatives quantitatives $x_1, \ldots, x_p$ au moyen d'un modèle linéaire :

$$y = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \epsilon$$

où ϵ est un terme d'erreur aléatoire de loi $N(0, \sigma^2)$.

2 Mise en œuvre sur un jeu de données simulées

- Simuler un jeux de données de taille $n = 100$ selon le modèle de régression linéaire multiple ci-dessous :

  $$y = 4 - 2x_1 + 3x_2 - 5x_3 + 0.8x_4 + 1.4x_5 + \epsilon$$

  où les x_k suivent la loi uniforme sur $[-5, +5]$ et ϵ suit la loi normale $N(0, \sigma^2 = 4)$.

  Après avoir construit les vecteurs `y`,`x1`, ..., `x5`, vous pouvez taper la commande ci-après pour construire le jeu de données correspondant :

  ```
  matYX <- data.frame(y,x1,x2,x3,x4,x5) # creation d'un "tableau de donnees"
  ```

- Il est possible d'obtenir un graphique de tous les nuages de points croisant deux variables du tableau des données :

  ```
  pairs(matYX)
  # ou de maniere equivalente
  plot(matYX)
  ```

- Pour faire la régression linéaire multiple, taper

  ```
  res <- lm(y~x1+x2+x3+x4+x5,data=matYX)
  summary(res)
  ```

  De même que pour la régression simple, il est possible de faire des graphiques intéressants (résidus, valeurs ajustées) :

  ```
  plot(res$fitted,res$residuals)
  abline(h=0)
  plot(res$fitted,y)
  abline(0,1)
  ```

  Il est aussi important de tester la normalité des résidus afin de s'assurer de la validité des tests d'hypothèses.

- Comparer les paramètres estimés avec les "vraies" paramètres du modèle simulé.
- Refaire ces manipulations avec un autre modèle (changer les paramètres β_k et le nombre p de variables explicatives).

3 Sélection de variables en régression linéaire multiple

Supposons que l'on travaille sur un modèle de régression de y sur les K régresseurs $x_1, \ldots, x_K$, les données étant contenues dans le jeu de données `matYX`.

3.1 Un critère pour choisir un "bon" modèle

Le critère utilisé par défaut dans R est le critère AIC (pour "An Information Criterion", proposé par Akaike, on parle aussi de critère d'Akaike).

Définition du critère du critère AIC (version simplifiée). La formule du critère AIC (sur lequel la méthodologie de sélection de variables est fondé) est la suivante :

$$AIC = n\ln\left(\frac{RSS^*}{n}\right) + 2(p^* + 1),$$

où n désigne la taille de l'échantillon, p^* correspond au nombre de variables explicatives considérées dans le modèle courant (i.e. celui pour lequel on est en train de calculer le AIC), et $RSS^* = \sum_{i=1}^{n}(y_i - \hat{y}_i^*)^2$ est la somme des carrés des résidus du modèle courant (RSS pour *Residual Sum of Squares* en anglais). Notons que la quantité $\frac{RSS^*}{n}$ correspond à l'estimation du paramètre σ^2 (variance du terme d'erreur ϵ du modèle) par la méthode du maximum de vraisemblance.

Rapide commentaire sur le critère AIC. La "philosophie" de ce critère est de trouver un compromis entre le modèle le plus simple possible (i.e. contenant le moins de variables explicatives) et qui ajustent le mieux les données (i.e. minimisant la somme du carrés des écarts entre les vraies valeurs et les valeurs ajustées par le modèle). Ce critère étant une fonction croissante de ces deux grandeurs p^* et RSS^*, le "meilleur" modèle (au sens du critère AIC) sera alors le modèle ayant la valeur du AIC la plus faible.

Remarque importante. Dans la littérature statistique, il existe de nombreux autres critères (BIC, C_p de Mallows, ...) pour faire de la sélection de variables et du choix de modèle. De nombreux packages R sont disponibles pour mettre en œuvre les différentes approches existantes.

3.2 Elimination d'une variable du modèle

On part d'un modèle "complet" et on cherche à supprimer (si nécessaire) la variable la moins "utile" au sens du critère AIC. Le modèle "complet" est estimé par les commandes

```
res <- lm(y~x1+x2+...+xK,data=matYX)
summary(res)
```

La fonction **drop1** retourne les effets sur le modèle de la suppression de chacune des K variables explicatives.

```
drop1(res)
```

Le critère utilisé est le critère AIC.

Interprétation rapide des sorties de la fonction "drop1". La ligne `<none>` correspond au modèle "complet" avec les K variables explicatives. Chaque ligne x_k correspond aux modèle dans lesquel on a supprimé la variable x_k. Si la plus faible des valeurs du critère AIC est celle d'une ligne `xk`, la variable correspondante peut être élimée du modèle. Il faut alors estimer le nouveau modèle de régression avec les $K-1$ variables explicatives restantes. Si le AIC le plus faible est celui de la ligne `<none>`, il ne faut pas supprimer de variables explicatives du modèle initial.

3.3 Ajout d'une variable explicative dans le modèle.

Supposons que l'on travaille sur un modèle de régression de y sur les k régresseurs $x_1, \ldots, x_k$. Il reste à notre disposition d'autres variables explicatives $x_{k+1}, \ldots, x_K$ qui peuvent éventuellement entrer dans le modèle.
Le modèle initial a été estimé par le biais des lignes de code suivantes :

```
res <- lm(y~x1+x2+...+xk,data=matYX)
summary(res)
```

La fonction **add1** retourne les effets sur le modèle initial de l'ajout d'une des variables explicatives $x_{k+1}, \ldots, x_K$.

```
add1(res,~x1+x2+...+xk+...+xK)
```

Le critère utilisé est le critère AIC.

Interprétation rapide des sorties de la fonction "add1". La ligne `<none>` correspond au modèle "initial" avec les k variables explicatives. Chaque ligne x_j correspond aux modèle dans lesquel on a ajouté aux k variables explicatives initiales la variable explicatives x_j (pour $j = k+1, \ldots, K$). Si la plus faible des valeurs du critère AIC est celle d'une ligne `xj`, la variable correspondante peut être ajoutée du modèle. Il faut alors estimer le nouveau modèle de régression avec les $k+1$ variables explicatives. Si le AIC le plus faible est celui de la ligne `<none>`, il ne faut pas ajouter de variables explicatives au modèle initial.

Remarque sur le choix de la première variable explicative du modèle. On peut regarder les corrélations linéaires entre y et chacune des K variables explicatives. Pour le modèle de régression linéaire simple (une seule variable explicative), on choisit celle qui est la plus corrélée avec y. Pour calculer les corrélations, vous pouvez taper les commandes suivantes :

```
cor(matYX)  # calcul de la matrice des correlations lineaires des donnees
cor(matYX[,1],matYX[,-1])  # calcul uniquement des corr.  entre les xk et y
```

On peut aussi utiliser la commande `lm`. Pour cela, taper les commandes suivantes :

```
res <- lm(y~1,data=matYX)   # on part d'un modele sans variable explicative
add1(res,~x1+x2+...+xK)  # on choisit la variable qui permet d'obtenir le "Cp" le plus faible.
```

3.4 Stratégies pour choisir le "meilleur" modèle de régression linéaire multiple avec R

On peut utiliser différents critères pour rechercher et définir le "meilleur" modèle. Dans la suite, nous proposons tout d'abord de simplifier le modèle de régression linéaire en se basant sur les résultats des tests de Student de nullité des paramètres associés aux régresseurs. Nous décrivons ensuite des techniques sélection de modèles en se basant sur le critère AIC : sélection ascendante, séléction descendante et sélection pas à pas.

• **Utilisation des tests de Student de nullité des paramètres associés aux régresseurs.** La démarche à suivre est ici de partir d'un modèle complet, c'est à dire contenant la totalité des p variables explicatives disponibles. L'estimation de ce modèle complet permet d'obtenir les p-values des tests de Student de nullité des paramètres associés aux régresseurs. On sort alors du modèle la variable la moins significative, c'est à dire celle qui a la p-value la plus élevée. On estime alors le nouveau modèle avec $p-1$ variables explicatives et on regarde à nouveau les p-values des tests de Student de de nullité des paramètres associés aux régresseurs. On sort alors du modèle la variable la moins significative. On réitère cette étape tant qu'il reste dans le modèle des variables non significative, c'est à dire dont la p-value est supérieur au risque de première espèce que l'utilisateur s'est fixé, par exmple $\alpha = 5\%$.

Remarque : N'oublier pas de contrôler à chaque étape, l'évolution de critères mesurant la bonne qualité du modèle estimé (R^2 ajusté, écart-type estimé du terme d'erreur, ...). Cela permet de vérifier si la suppression d'une variable "améliore" ou "détériore" le modèle.

• **Sélection ascendante (forward selection).** La philosophie des cette méthode est de partir d'un modèle ne contenant pas de variables explicatives, puis de faire rentrer à chaque étape une variable dans le modèle, jusqu'à ce que l'ajout d'une variable supplémentaire dans le modèle n'apporte plus rien au modèle. Le critère utilisé peut être le critère AIC. On peut alors utiliser à chaque étape la fonction **add1**.

• **Sélection descendante (backward selection).** La philosophie de cette méthode est de partir d'un modèle contenant la totalité des variables explicatives disponibles, puis de faire sortir à chaque étape une variable du modèle, jusqu'à ce que la suppression d'une variable du modèle détériore sensiblement le modèle. Le critère utilisé peut être le critère AIC. On peut alors utiliser à chaque étape la fonction **drop1**.

• **Sélection pas à pas (stepwise selection).** La philosophie de cette méthode est celle d'un mélange de la sélection ascendante et de la sélection descendante. Ainsi, à chaque étape, une étape de sélection ascendante est effectuée, ainsi qu'une étape de sélection descendante. La procédure s'arrête lorsque l'ajout et la suppression d'une variable dans le modèle ont des conséquences néfastes le modèle au vu du critère utilisé, ici encore le critère AIC par exemple peut être utilisé. Pour cette sélection pas à pas, il existe une fonction particulière : la fonction **step**.

Si on part d'un modèle complet, on fait de la régression pas à pas "descendante". Si on part d'un modèle sans variable explicative ou avec peu de variables explicatives, on fait de la régression pas à pas "ascendante".

Exemple de régression pas à pas ascendante : utilisation de la fonction **step**.

```
res <- lm(y~1,data=matYX) # modele initial sans variable explicative, les var. etant dans "matYX"
step(res,~x1+...+xp)   # avec x1,...,xp,  p colonnes de "matYX"
```

Il existe une option utile de la fonction **step**, il s'agit de l'option *trace=F* qui permet de ne récupérer que le modèle final sans détailler les différentes étapes.

Exemple de régression pas à pas descendante : utilisation de la fonction **step**.

```
res <- lm(y~x1+...+xp,data=matYX) # modele initial complet
step(res)
```

3.5 Travail à effectuer : sélection de variables sur des simulations

- Simuler un jeu de données provenant d'un modèle de régression linéaire multiple contenant un nombre $p \geq 10$ de variables explicatives et dans lequel on choisit un nombre $K \leq 3$ de coefficients β_k nuls.

  Utiliser les différentes méthodes de sélection de variables pour obtenir le meilleur modèle pour ce jeu de données simulées. Comparer les modèles obtenus. Commenter.

- Refaire les différentes manipulations sur de nouvelles données simulées avec un nouveau modèle de régression linéaire multiple.

4 Prédiction dans un modèle de régression linéaire multiple

Pour faire de la prédiction dans un modèle de régression linéaire, on peut utiliser la fonction **predict**.

Supposons que l'on dispose de $p+1$ vecteurs `y, x1, ..., xp` de taille n contenant respectivement les valeurs de la variable à expliquer et des p variables explicatives, stockés dans le jeu de données `matYX`. On désire faire de la prévision sur m valeurs des variables explicatives contenues dans les vecteurs `nouveaux1, ..., nouveauxp` de taille m. Pour cela, vous pouvez lancer les lignes de commandes suivantes :

```
matYX <- data.frame(y,x1,...,xp)
res <- lm(y~x1+...+xp,data=matYX)
predict(res,data.frame(x1=nouveaux1,...,xp=nouveauxp))
```

Les différentes options disponibles pour cette fonction **predict** (que l'on a décrites dans le TP portant sur la régression linéaire simple) sont utilisable dans ce cadre de régression linéaire multiple.

5 Exercices

5.1 Etude du jeu de données réelles "ozone"

La problématique et les données disponibles. Nous allons dans cette partie étudier le jeu de données "ozone" (contenu dans le fichier `ozone.txt`). Il s'agit de données concernant la pollution de l'air, sujet qui constitue une des préoccupations majeures de santé publique. De nombreuses études ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques présents dans l'air comme le dioxide de souffre (SO_2), le dioxyde d'azote (NO_2), l'ozone (O_3) ou encore des particules de poussières fines.

De nombreuses associations de surveillance de la qualité de l'air existent sur tout le territoire français et mesurent la concentration des polluants présents dans l'air. Elles enregistrent aussi simultanément les conditions météorologiques comme la température, la nébuloisité, le vent, ...

Les données contenues dans ce fichier concernent $n = 112$ relevés faits durant l'été 2001 Rennes.

Dans le fichier, la variable à expliquer est le maximum journalier de la concentration en ozone (en $\mu g/m^3$), variable `max03`, en fonction des autres variables quantitatives disponibles : des variables de température `T9, T12, T15`, des variables de nébulosité `Ne9, Ne12, Ne15`, des variables de vent `Vx9, Vx12, Vx15`, et aussi de la mesure du maximum de la concentration en ozone de la veille `max03v`.

Travail à réaliser. A partir de ces données, proposer le meilleur modèle de régression linéaire multiple pour expliquer le maximum journalier de la concentration en ozone. Justifier votre démarche.

5.2 Etude sur des données réelles "station"

La direction marketing d'un distributeur d'essence souhaite établir un modèle expliquant les ventes de ses stations services situées dans les grands centres urbains. Le tableau de données précise, pour 45 stations de ce type, les informations suivantes :

- les ventes de la station exprimées en milliers de litres (variable `ventes`),
- le nombre de pompes de la station (variable `nbpompes`),
- le nombre de concurrents dans la zone desservie par la station (variable `nbconc`),
- le trafic quotidien exprimé en milliers de voitures (variable `trafic`).

Les données sont contenues dans le fichier `station.txt`. L'objectif de l'étude est donc d'essayer de modéliser les ventes d'une station en fonction des autres variables disponibles.

Le statisticien chargé de l'étude va mettre en œuvre toutes les régressions linéaires (simples ou multiples) à une, deux et trois variables explicatives, pour modéliser le nombre de ventes.

1. Expliquer pourquoi le modèle à trois variables ne peut être retenu. Quelle variable faut-il éliminer de ce modèle pour obtenir un modèle à deux variables qui semble mieux convenir ? Justifier votre choix.
2. a) Faire une analyse des indices de qualité du modèle à deux variables obtenu à la question précédente et en tirer une conclusion quant à sa validité statistique.

   b) Faire une interprétation des cœfficients de ce modèle. Que pensez-vous de la validité économique du modèle ? Permet-il de mieux comprendre la réalité ?

   c) Préciser ce que l'on prédirait comme nombre de ventes étant donné un nouveau couple de variables explicatives.
3. Justifier pourquoi le modèle à deux variables défini à la première question et étudié à la deuxième est-il préférable aux modèles à une seule variable ?