TD-TP Régression linéaire multiple

Ce qui change par rapport à la régression linéaire simple

Modèle M_{123} : $Y_i = \mu + a\ x_{1,i} + b\ x_{2,i} + c\ x_{3,i} + \epsilon_i,\ \ i = 1, \ldots, n$
Fonction `lm()` : res=lm($y \sim x_1 + x_2 + x_3$, data) ou plus simplement res=lm($y \sim .$, data) si on veut le modèle avec toutes les variables explicatives possibles (i.e toutes les variables sauf y) du fichier data.
Résultats :
La commande `summary(res)` donne une liste contenant entre autres :

— un tableau :

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	$\hat{\mu}$	$\hat{\sigma}_{\hat{\mu}}$	statistique $\frac{\hat{\mu}-0}{\hat{\sigma}_{\hat{\mu}}}$	p value du test de $H_0 : \mu = 0$
x_1	$\hat{a}$	$\hat{\sigma}_{\hat{a}}$	statistique $\frac{\hat{a}-0}{\hat{\sigma}_{\hat{a}}}$	p value du test de $H_0 : a = 0$ (*)
x_2	$\hat{b}$	$\hat{\sigma}_{\hat{b}}$	statistique $\frac{\hat{b}-0}{\hat{\sigma}_{\hat{b}}}$	p value du test de $H_0 : b = 0$
x_3	$\hat{c}$	$\hat{\sigma}_{\hat{c}}$	statistique $\frac{\hat{c}-0}{\hat{\sigma}_{\hat{c}}}$	p value du test de $H_0 : c = 0$

(*) i.e test de $H_0 : Y_i = \mu + b\ x_{2,i} + c\ x_{3,i} + \epsilon_i$ contre $H_1 : Y_i = \mu + a\ x_{1,i} + b\ x_{2,i} + c\ x_{3,i} + \epsilon_i$

— la statistique "F-statistic" du test de Fisher, les ddl("DF") et la p-value associés :
Test de H_0 : modèle $M_0 : y_i = \mu + \epsilon_i$
contre H_1 : modèle $M_{123} : y_i = \mu + a\ x_{1,i} + b\ x_{2,i} + c\ x_{3,i} + \epsilon_i$

$$\text{F-statistic} = \frac{(\text{SCR}_{M_0} - \text{SCR}_{M_{123}})/(\dim M_{123} - \dim M_0)}{\text{SCR}_{M_{123}}/(n - \dim M_{123})} = \frac{||\hat{Y}_{M_{123}} - \hat{Y}_{M_0}||^2/(\dim M_{123} - \dim M_0)}{||Y - \hat{Y}_{M_{123}}||^2/(n - \dim M_{123})}$$

avec $(\hat{Y}_{M_{123}})_i = \hat{\mu} + \hat{a}\ x_{1,i} + \hat{b}\ x_{2,i} + \hat{c}\ x_{3,i}$ et $(\hat{Y}_{M_0})_i = \hat{\mu}$

La commande `anova(res)` permet en plus la sortie de la table d'analyse de variance.

Modèle $M_0 : y_i = \mu + \epsilon_i$
Modèle $M_1 : y_i = \mu + a\ x_{1,i} + \epsilon_i$
Modèle $M_{12} : y_i = \mu + a\ x_{1,i} + b\ x_{2,i} + \epsilon_i$
Modèle $M_{123} : y_i = \mu + a\ x_{1,i} + b\ x_{2,i} + c\ x_{3,i} + \epsilon_i$

	ddl	Somme carrés	Moyenne des carrés	F value
x_1	dim $M_1 - 1$	$\|\|\hat{Y}_{M_1} - \hat{Y}_{M_0}\|\|^2$	$N_1 = \|\|\hat{Y}_{M_1} - \hat{Y}_{M_0}\|\|^2/(\dim M_1 - 1)$	$\frac{N_1}{D}$
x_2	dim M_{12} - dim M_1	$\|\|\hat{Y}_{M_{12}} - \hat{Y}_{M_1}\|\|^2$	$N_2 = \|\|\hat{Y}_{M_{12}} - \hat{Y}_{M_1}\|\|^2/(\dim M_{12} - \dim M_1)$	$\frac{N_2}{D}$
x_3	dim M_{123} - dim M_{12}	$\|\|\hat{Y}_{M_{123}} - \hat{Y}_{M_{12}}\|\|^2$	$N_3 = \|\|\hat{Y}_{M_{123}} - \hat{Y}_{M_{12}}\|\|^2/(\dim M_{123} - \dim M_{12})$	$\frac{N_3}{D}$
Res	n-dim M_{123}	$\|\|Y - \hat{Y}_{M_{123}}\|\|^2$	$D = \|\|Y - \hat{Y}_{M_{123}}\|\|^2/(\text{n} - \dim M_{123})$	

— La première ligne de ce tableau permet de tester le modèle M_0 contre le modèle M_1 avec une loi de Fisher de ddl $(\dim M_1 - 1)$ et $(n - \dim M_{123})$. Ce n'est pas tout à fait la statistique de Fisher vue en cours. En cours : $F = \frac{||\hat{Y}_{M_1} - \hat{Y}_{M_0}||^2/(\dim M_1 - 1)}{||Y - \hat{Y}_{M_1}||^2/(n - \dim M_1)} \sim \mathcal{F}(\dim M_1 - 1, n - \dim M_1)$
— La deuxième ligne de ce tableau permet de tester le modèle M_1 contre le modèle M_{12}, et la troisième ligne de tester le modèle M_{12} contre le modèle M_{123}.

Remarque : La deuxième ligne du tableau `anova(res)` permet donc de tester l'apport de la variable x_2 une fois que x_1 est accepté. Si vous faites :

resBIS=lm($y \sim x_2 + x_1 + x_3$,data)

la deuxième ligne du tableau `anova(resBIS)` permettra de tester l'apport de la variable x_1 une fois que x_2 est accepté. On n'obtiendra pas forcément les mêmes résultats : En règle géneral, tester d'abord la significativité de x_1 puis de x_2 n'est pas équivalent à tester la significativité de x_2 puis de x_1.

Colinéarité entre variables : Dans le cas où des variables explicatives ont une colinéarité excessive, alors le problème est dit mal conditionné, au sens où la matrice X ne va pas être de plein rang. On va donc chercher à vérifier qu'il n'y a pas de colinéarité entre les variables explicatives. Pour cela, on considére la régression d'une variables explicative x^j en fonction des autres variables explicatives et on calcule le coefficient de détermination R_j^2 associé à cette régression. Si R_j^2 est proche de 0, alors la variable explicative x^j n'est pas fonction linéaire des autres variables. Si ce R_j^2 est proche de 1, alors elle est fonction linéaire des autres varaiables et on a un problème de colinéarité.

On définit les coefficients de " variance inflation factor" (VIF) par $VIF_j = \frac{1}{1-R_j^2} \quad \forall j$

Si $VIF_j > 10$, on considère qu'il y a un problème de colinéarité de la variable x^j, et on enlève cette variable explicative.

Comparaison de deux modèles emboitées : Pour comparer deux modèles emboitées M_a et M_b avec M_b plus géneral, on peut utiliser la commande anova(res$M_a, resM_b$), qui calcule la statistique de Fisher :

$$F = \frac{\left(SCR_{M_a} - SCR_{M_b}\right)/(\dim M_b - \dim M_a)}{SCR_{M_b}/(n - \dim M_b)}$$

Application

Etude de l'adaptation d'une variéte de moutarde à la sécheresse : $n = 31$ pieds de moutarde sont soumis à un stress hydrique de 34 jours après repiquage. On veut étudier le nombre de racines courtes tubérisées RC (qui est un mécanisme d'adaptation à la sécheresse de la moutarde) en fonction de 5 variables mesurées : X_1 la longueur de la tige, X_2 le potentiel hydrique foliaire, X_3 le poids de matière sèche des racines, X_4 le poids de matière sèche des parties aériennes et X_5 le nombre de feuilles. On considère ce modèle de régression :

$$RC_i = \mu + \sum_{k=1}^{5} \beta_k X_{k,i} + \varepsilon_i, \ i = 1, \ldots, n. \tag{1}$$

1. Récupérer le fichier de données et l'ouvrir sous R. en tapant :

   ```
   data=read.table("moutarde.txt",header=TRUE)
   attach(data)
   ```

2. (sur papier) Mettre le modèle (1) sous forme matricielle. Quelle hypothèse usuelle sur la matrice X est nécessaire pour calculer de manière unique l'estimateur des moindres carrés (EMC) de θ.
   Vérifier qu'il n'y a pas de problème de colinéarité entre les variables :

   ```
   library(car)
   vif(res)
   ```

3. Que doivent vérifier les ε_i ? Effectuez l'analyse des résidus pour valider les hypothèses du modèle. Qu'en concluez-vous ?
4. On suggère les transformations suivantes pour la variable RC : logarithme - racine carree - inverse. Laquelle de ces transformations choisissez-vous et pourquoi ?
5. Pour la transformation de RC que vous aurez choisie, vérifier la normalité des résidus avec le test de Shapiro-Wilk (H_0 : normalité) avec

   ```
   shapiro.test(res$residuals)
   ```

   Dans la suite on va considérer le modèle (1) avec la transformation de RC que vous aurez choisie. Vous pouvez par exemple noter $y = log(RC)$, et expliquer y à partir de $X_1, \ldots, X_5$.
6. (sur papier) Rappeler l'expression de l'estimateur $\hat{\theta}$ des moindres carrés de θ.
   Application avec les sorties de R : donner les estimations des paramètres $\beta_k, \quad k = 1, \ldots, 5$

7. On souhaite tester $H_0 : \beta_k = 0, \forall k = 1, \ldots, 5$ contre $H_1 : \exists k$ tel que $\beta_k \neq 0$.
   (a) Donner les modèles associés aux hypothèses H_0 et H_1 de ce test.
   (b) On s'intéresse au modèle sous H_0 : le mettre sous forme matricielle ; donner la dimension de ce modèle ; donner l'expression des valeurs prédites $\hat{Y}_i$ par ce modèle en fonction des paramètres estimés ; en déduire la somme des carrés résiduelles sous ce modèle.
   (c) Même question pour le modèle sous H_1.
   (d) Effectuer le test au niveau 5% : donner la statistique de test et conclure.
8. On souhaite maintenant tester l'effet de chaque variable explicative au niveau 5%, i.e tester pour chacune des 5 variables si on peut se dispenser de celle-ci. (Par exemple pour $x1$, test de $Y_i = \mu + \sum_{k=2}^{5} \beta_k X_{k,i} + \varepsilon_i$ contre $H_1 : Y_i = \mu + \sum_{k=1}^{5} \beta_k X_{k,i} + \varepsilon_i$.
   Donner la statistique de test appropriée et sa loi. Conclure.
   Retrouver le résultats de ces tests en utilisant un intervalle de confiance.
9. Test unilatéral : tester $\beta_3 = 0$ contre $\beta_3 > 0$.
10. Faire un test unilatéral sur la paramètre β_4.
11. L'expérimentateur aimerait, pour une expérience ultérieure se dispenser de la mesure du potentiel hydrique, X_2, qui est lourde à réaliser, du comptage des feuilles, X_5, ainsi que de la pesée de la partie aérienne X_4.
    Donner les modèles (et leur dimensions) associés aux hypothèses H_0 et H_1 de ce test. Donner les expressions des valeurs prédites par chacun de ces deux modèles. Donner la statistique de test et conclure : le sous-modèle convient-il ?
12. Quel sous modèle convient ?

Choix de modèles : algorithme de sélection

Lorsque l'on dispose de beaucoup de variables explicatives potentielles, toutes ne sont pas pertinentes. On souhaiterait une procédure/algorithme pour choisir un "bon" modèle.

— Il s'agit donc de définir ce qu'on entend par "bon" modèle, ie définir un critère.

— Puis de rechercher le meilleur modèle au sens de ce critère. Soit avec une recherche exhaustive, i.e en comparant les 2^p modèles possibles lorque l'on dispose de p variables explicatives possibles. En pratique, dès que p est un peu grand, cela devient très couteux en temps de calcul. On va alors se pencher vers des algorithmes pas à pas.

Algorithmes pas à pas :

- **Forward** : On part du modèle de régression le plus simple (dans lequel on ne met que l'effet moyen μ). Puis on ajoute la variable jugée la plus significative (selon un critère choisi) et on itère la procédure.
- **Backward** : On part du modèle de régression multiple avec toutes les variables. Puis on retire la variable jugée la moins significative (selon un critère choisi) et on itère la procédure.
- **Stepwise** : Combinaison des deux méthodes : on peut retirer ou ajouter des variables à chaque étape. En effet, une variable considérée comme la plus significative à une étape de l'algorithme peut à une étape ultérieure devenir non significative, en raison de corrélations avec d'autres variables introduites après coup dans le modèle.

Choix du critère : recherche d'un bon modèle explicatif :
Un bon modèle explicatif est un modèle qui permet d'approcher la réalité, ie le "vrai" modèle supposé exister sur les données. Ce "vrai modèle" est éventuellement basé sur une théorie physique, économique, bilogique... On veut donc savoir quelles variables ont une influence sur la réponse, pour pouvoir les interpréter, comprendre le phénomène. Le rôle de chaque variable explicative est important.

Critères possibles : ce sont des critères dits d'ajustement.

— La statistique du test de Fisher qui permet de tester le caractère significatif d'une ou plusieurs variables sur la réponse.

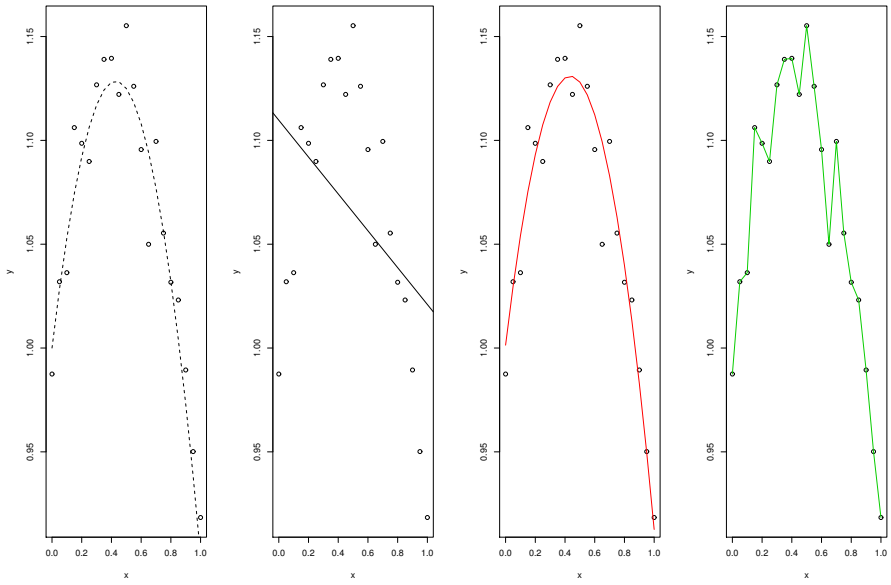
— le coefficient de détermination R^2 qui mesure la qualité de l'ajustement d'un modèle sur les données. Comme ce R^2 augmente avec le nombre de paramètres, si on veut comparer des modèles ne comportant pas le même nombre de variables explicatives, on préfère regarder le R^2 ajusté qui prend en compte la dimension p du modèle :

$$R^2 adjust = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)}$$

En pratique, il est souvent difficile de trouver un modèle explicatif facilement interprétable, en particulier s'il existe de la colinéarité entre les variables.

Choix du critère : recherche d'un bon modèle prédictif - Estimation versus Apprentissage :

On peut préférer un bon modèle prédictif, ie un modèle qui prédira de nouvelles observations avec une bonne précision. C'est le cadre de l'apprentissage, où l'objectif est essentiellement la prévision. Le meilleur modèle dans ce sens, n'est pas nécessairement celui qui ajusterait le mieux le "vrai" modèle.



(x_i, y_i) simulés selon un modèle polynomiale : $y_i = f(x_i) + \epsilon_i$ avec f fonction polynomiale de degré 2 donnée représentée en pointillé sur le graphe de gauche. Les 3 graphes suivants correspondent à :
Ajustement (1) par une droite, (2) par un polynome de degré2, et (3) par un polynome de degré 20.

(1) : modèle d'ajustement trop simple
(3) : modèle qui s'ajuste trop bien aux données. Pourquoi trop bien ? le modèle capture aussi le bruit des données même s'il n'est pas représentatif du phénomène. Ce modèle est trop variable (si on retire un point de nos données, on obtient un modèle différent). Ce modèle prédira mal de nouvelles données provenant du même phénomène.

Le choix d'un bon modèle prédictif est basé sur l'erreur de prédiction. Supposons que l'on dipose d'un jeu de données $(x_i, y_i)_{i=1,\dots,n}$, appelé jeu d'apprentissage (training set en anglais). Le "vrai" modèle sur ces données est du type $y = f(x) + \epsilon$ avec f une fonction inconnue (par exemple f est polynomiale d'odre 2 $f(x) = a + bx + cx^2$). Avec le jeu d'apprentissage, on estime notre fonction f inconnue selon un certain modèle (par exemple avec un modèle de regression linéaire simple $\hat{f}(x) = \hat{a} + \hat{b}x$).
Supposons que l'on dispose d'une nouvelle donnée (x_{n+1}, Y_{n+1}). On appelle en apprentissage (x_{n+1}, Y_{n+1}) le jeu de données test : ce sont des données indépendantes des $(x_i, y_i)_{i=1,\dots,n}$, et issues du même phénomène, ie du même "vrai" modèle f. On se demande comment la fonction $\hat{f}$ prédite avec le jeu d'apprentissage, va prédire la nouvelle observation. Pour le mesurer, on définit l'erreur moyenne de prédiction par :

$$\mathbb{E}\left[\left(Y_{n+1} - \hat{f}(x_{n+1})\right)^2\right]$$

On peut montrer que cette erreur moyenne de prédiction se décompose en 3 termes :

$$\mathbb{E}\left[\left(Y_{n+1} - \hat{f}(x_{n+1})\right)^2\right] = \sigma^2 + \text{Biais}^2(\hat{f}(x_{n+1})) + \text{Var}(\hat{f}(x_{n+1}))$$

où

1. $\sigma^2 = \text{Var}(\epsilon)$ est la variable résiduelle du modèle. L'erreur de prédiction est forcément supérieure à σ^2 : c'est l'erreur dite inévitable dûe au bruit des données.
2. $\text{Biais}(\hat{f}(x_{n+1})) = \mathbb{E}[\hat{f}(x_{n+1})] - f(x_{n+1})$ représente l'erreur lorsque l'on approxime le vrai modèle f par le modèle $\hat{f}$. Ainsi si f est polynomiale d'ordre 2 et que l'on choisit $\hat{f}$ selon un modèle de régression linéaire simple, on a une erreur de biais. Cette erreur de biais est grande si f correspond à un modèle "compliqué" (avec beaucoup de paramètres) et que l'on choisit $\hat{f}$ un modèle trop "simple" (avec peu de paramètres).
3. La variance $\text{Var}(\hat{f}(x_{n+1})) = \mathbb{E}\left[\left(\hat{f}(x_{n+1}) - \mathbb{E}(\hat{f}(x_{n+1}))\right)^2\right]$ représente à quel point la fonction $\hat{f}$ est variable. Autrement dit, si je retire un point de mon échantillon d'appentissage, est-ce-que mon modèle $\hat{f}$ reste sensiblement le même ? On parle d'erreur de variance. L'erreur de variance augmente avec le nombre de paramètres à estimer dans le modèle $\hat{f}$.

Choisir un modèle avec une faible erreur de prédiction revient donc à faire un compromis biais-variance, puisque que les deux varient en sens inverse : un modèle "simple" avec peu de paramètres aura un grand biais et une faible variance (on parle de sous-apprentissage, c'est le cas (1) de la figure), et inversement un modèle "compliqué" avec beaucoup de paramètres aura un faible biais mais une grande variance (on parle de sur-apprentissage, c'est le cas (3) de la figure).
Le principe de la sélection de modèle est de rechercher un modèle qui fait un compromis biais-variance, ie un compromis entre qualité de l'approximation et complexité du modèle. On considère des critères de vraisemblance pénalisés du type :

$$crit(m) = -2log(L(m)) + pen(m)$$

où $log(L(m))$ est la log-vraisemblance du modèle m (évaluée en l'EMV des paramètres) et $pen(m)$ un terme qui va pénaliser les modèles de trop grande dimension. On cherche alors le modèle m qui minimise $crit(m)$.
Les deux critères les plus usuels sont les suivants :

Le critère d'Akaike AIC

$$AIC(m) = -2log(L(m)) + 2|m|$$

Le critère Bayesian Information criterion BIC

$$BIC(m) = -2log(L(m)) + |m|log(n)$$

Le terme de pénalité $pen(m)$ dépend donc de la dimension $|m|$ du modèle m. Lorsque $n \geq 8$ alors $log(n) \geq 2$ et la pénalité du critère BIC est plus lourde que celle de l'AIC ce qui fait que les modèles choisis par BIC auront moins de variables explicatives. En pratique, on calcule le critère AIC (ou BIC) pour plusieurs modèles (par exemple selon un algorithme pas à pas) et on choisit le modèle qui possède l'AIC (ou le BIC) le plus faible.

Remarque. En régression linéaire multiple, sous hypothèse de normalité des résidus, $log(L(m))$ s'écrit

$$log(L(m)) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\hat{\sigma}_m^2) - \frac{1}{2\hat{\sigma}_m^2}||Y - X_m\hat{\theta}_m||^2$$

avec X_m la matrice obtenue lorsqu'on met le modèle m sous forme matricielle (par exemple m est le modèle de régression simple avec la première variable explicative possible), et $\hat{\sigma}_m^2$ et $\hat{\theta}_m$ les EMV de σ^2 et de θ dans le modèle m. A constance près $-2log(L(m))$ s'écrit $nlog(\frac{SCR(m)}{n})$ car $\hat{\sigma}_m^2 = \frac{SCR(m)}{n}$. Donc maximiser la vraisemblance revient à minimiser la SCR.

Procedure `step` sous R :

- **Forward :**

```
res0=lm(y ~ 1,data)   # Modele contenant uniquement l effet moyen mu
resT=lm(y~x1+x2+x3+x4+x5 ,data)    # Modele contenant toutes les variables explicatives

modele_Forward=step(res0,scope=formula(resT),direction="forward")
# Voici ce que l on obtient :
# Le signe + indique qu on tente d ajouter la variable en question au
modele. La valeur de AIC est celle du modele contenant la variable en question.
# <none> réfère au modèle précédent.

Start:  AIC=27.98                    # point de depart : aucune variable (res0)
```

```
y ~ 1
       Df Sum of Sq    RSS     AIC    # Étape 1
+ x3    1    62.448  9.221 -33.588    # AIC le plus bas, < AIC(none)
+ x2    1    46.261 25.408  -2.167
+ x1    1    37.983 33.685   6.575
+ x5    1    30.456 41.213  12.827
+ x4    1     8.868 62.801  25.885
<none>              71.669  27.980    # AIC du modele avec 0 var.
Step:  AIC=-33.59
y ~ x3                                # Premier modele : 1 variable explicative
       Df Sum of Sq    RSS     AIC    # Étape 2
+ x4    1   1.56202 7.6588 -37.342    # AIC le plus bas, < AIC(none)
+ x2    1   1.05577 8.1651 -35.358
<none>              9.2209 -33.588    # AIC du modele avec 1 variable.
+ x5    1   0.13403 9.0868 -32.042
+ x1    1   0.08304 9.1378 -31.869
Step:  AIC=-37.34
y ~ x3 + x4                            # Second modele : 2 variables explicatives
       Df Sum of Sq    RSS     AIC     # Étape 3
<none>              7.6588 -37.342     # AIC du modele avec 2 variables.
+ x1    1   0.40105 7.2578 -37.009     # AIC > AI(none) donc variable non selectionnee
+ x2    1   0.37345 7.2854 -36.892
+ x5    1   0.12136 7.5375 -35.837
# on retient donc le modele avec les variables x3 etx4.

# vous pouvez taper summary(modele$_$Forward) pour étudier ce modèle
```

- **Backward :**
  ```
  resT=lm(y~ x1+x2+x3+x4+x5,data)    # Modèle contenant toutes les variables explicatives
  modele_Backward=step(resT,direction="backward")
  ```
- **Stepwise :**
  ```
  modele_Stepwise=step(res0,scope=formula(resT),direction="both")
  ```

Comparer les résultats obtenus avec les procédures Forward, Backward, Stepwise.

Remarque : Toutes ces procédures ne mènent pas forcément à la même solution.

— La procédure forward evite de travailler avec plus de variables que nécessaire.

— La procédure backward est satisfaisante pour l'utilisateur préférant avoir toutes les variables possibles afin de ne rien ignorer.

— Mais pour ces deux procédures, une fois qu'une variable a été ajoutée (ou éliminée), on ne peut plus la retirer (ou l'ajouter).

Dans la pratique, les procédures backward et stepwise sont les plus utilisées.

Remarque : pour une recherche exhaustive, i.e pour comparer les 2^p modèles possibles (lorsque p n'est pas trop grand) : sous R, package leaps.

Exercice d'entrainement

Utiliser les données *ozone* et expliquer la variable $maxO3$ en fonction des autres variables quantitatives (si vous utilisez aussi les variables explicatives, alors ce n'est plus de la régression linéaire multiple, mais un modèle de covariance). Vous pourrez :

1. Etudier les colinéarités entre variables explicatives
2. Etudier les résidus : appliquer une transformation log pour stabiliser la variance ? retirer des données (une à la fois) qui vous semblent atypiques ?
   Remarque : trouver un modèle qui vous convient n'est pas si facile. On peut toujours affiner et retirer des données, mais il faut s'arrêter lorsqu'un modèle vous semble convenable. (Vous n'aurez pas forcément tous le même modèle).
3. Etudier le modèle choisi : quelles variables semblent significatives ? Dans quelle sens une variable influe t-elle sur la concentration en ozone (test unilatéral) ? Faire de la prédiction avec la fonction *predict*.
4. Mettre en oeuvre une procédure de sélection de modèle pour proposer un modèle de dimension réduite.