Transcription de l'audio du cours de modélisation (F. Husson)

Diapositive 1 :

Bonjour,

Ce cours de modélisation se divise en 2 parties. Dans une première partie nous allons voir comment construire et sélectionner un modèle. Dans une seconde partie nous verrons comment interpréter les résultats d'un modèle et comment utiliser un modèle pour faire de la prévision.

Diapositive 2 :

Mais tout d'abord, qu'est-ce que la modélisation ? On peut résumer succinctement en disant que modéliser c'est d'une part comprendre, et d'autre part prévoir. Et qui a déjà fait de la modélisation ? A mon avis, chacun de nous ! Et même en construisant des modèles complexes … mais le plus souvent sans le savoir. En effet, on va prévoir un temps de trajet en fonction du jour de la semaine, de l'heure de départ, du fait qu'il pleuve ou non, que c'est un jour de départ en vacances, etc, etc. On peut aussi chercher à estimer le temps d'attente à un guichet en fonction du nombre de personnes, du nombre de guichets ouverts, etc. Estimer le prix de location d'un appartement en fonction de sa superficie, du nombre de chambres, de sa localisation, de son exposition, de la présence d'espaces verts, etc. Dans ces 3 premiers exemples, on cherche à estimer une variable quantitative. On peut aussi chercher à prévoir une variable qualitative, comme lorsqu'on décide comment s'habiller le matin, en prenant ou pas un vêtement de pluie. Pour prédire, on utilise là encore des informations comme la pluviométrie des jours précédents, la présence de nuages dans le ciel, la présence de vent, etc. Cette fois, la variable « va-t-il pleuvoir » que l'on cherche à prédire est qualitative puisque 2 réponses sont possibles oui/non. Ce type de modélisation ne sera pas abordée dans ce cours même si beaucoup d'idées peuvent être reprises puisqu'on essaie de comprendre l'effet de plusieurs variables sur une variable réponse, mais ici la variable réponse est qualitative.

Et comment faites-vous pour prévoir ? Vous listez toutes les variables qui potentiellement peuvent influer sur la variable réponse, c'est-à-dire la variable que vous cherchez à comprendre. Une fois que vous avez l'ensemble des variables, vous ignorez les variables qui sont négligeables et qui n'aident pas à mieux prévoir votre réponse, et vous essayez ensuite de quantifier au mieux l'effet des variables restantes qui ont été sélectionnées.

Et alors, est-ce que cette stratégie est bien raisonnable ? En fait, oui tout à fait. Et c'est exactement la démarche que l'on va adopter en modélisation. Mais alors, on peut se demander « à quoi servent les statistiques ? ». A faire cela avec rigueur, pour comprendre et prédire des variables qui mesurent des phénomènes parfois beaucoup plus complexes. Mais l'idée de cette démarche est à garder pour la modélisation.

Diapositive 3 :

Voici quelques exemples de modèles en agronomie ou alimentation, mais on trouve de tels modèles dans de très nombreux domaines comme l'économie, la démographie, la climatologie, etc.

Premier exemple : on cherche à prédire le temps de cuisson idéal en fonction de la composition et du poids d'un aliment, de la température du four, de l'humidité de l'air, …

Deuxième exemple : on veut prévoir la production de biogaz en fonction de la quantité de déchets agricoles ou alimentaires, des résidus de culture, d'ordures ménagères, etc.

3ème exemple : on veut prévoir la production électrique d'une éolienne en fonction de la vitesse du vent à 10m, à 80m, de la température à 2m, de la pression atmosphérique, de l'humidité relative à 2m, de la direction du vent

Dernier exemple, on veut optimiser le couple durée - température pour maximiser un rendement. On va alors construire un modèle du rendement en fonction du temps et de la température, et une fois le modèle connu, on pourra prédire le rendement pour n'importe quel couple … et choisir le couple qui optimise le rendement.

Dans tous ces exemples de modélisation, les objectifs sont de comprendre quelles variables influent sur une variable réponse quantitative, et également de prévoir les valeurs de la réponse pour de nouvelles conditions.

Diapositive 4 :

Nous allons illustrer ce cours avec un exemple sur la prévision de la qualité de l'air. Plus précisément, nous allons chercher à prédire le maximum d'ozone atteint lors d'une journée, et ce en fonction de diverses variables météorologiques comme la température à 9h, à 12h à 15h , la nébulosité à 9h, 12h 15h, la vitesse du vent calculé sur un axe ouest-est (si le vent vient de l'ouest la valeur sera négative, si le vent vient de l'est la valeur sera positive) toujours mesuré à 9h, 12h et 15h, la valeur du maximum d'ozone de la veille, la direction du vent (avec 4 modalités possibles : nord, est, sud, ouest) ou la pluviométrie considéré comme une variable qualitative à 2 modalités (sec ou pluvieux). L'objectif de cette étude est alors de construire un modèle pour comprendre d'une part ce qui peut influer sur le maximum d'ozone, et ensuite de construire un modèle qui permettra de prédire le maximum d'ozone pour de nouvelles données météo. Les lignes du jeu de données correspondent à des jours de mesure (à partir du 1er juin 2001), et nous avons 112 relevés à Rennes. Les données sont présentes dans ce fichier disponible sur internet. On peut importer le fichier dans le logiciel R en utilisant la fonction read.table, en précisant le lien vers le fichier, en précisant que la 1ère ligne du jeu de données contient le nom des variables avec header=TRUE, et en précisant que les variables textuelles sont en réalité des facteurs, ie.e. des variables qualitatives grâce à stringAsFactors=TRUE.

Diapositive 5 :

Avant de se lancer dans la modélisation, il est toujours indispensable de visualiser les données. Visualiser les données permet d'une part de s'assurer que nous n'avons pas de données aberrantes, et d'autre part cela permet d'avoir une idée des liaisons entre les variables, et notamment entre la variable réponse et les autres variables. Ce graphique représente le maximum d'ozone en fonction de la température à 9h. Il permet de voir par exemple si la liaison est plutôt linéaire (les points sont plus ou moins autour d'une droite), ou si la liaison n'est pas linéaire (par exemple une liaison quadratique ou exponentielle) ou s'il n'y a pas de liaison entre ces 2 variables. Evidemment, si on visualise une liaison plutôt quadratique, on ne pourra pas bien prédire avec des effets linéaires uniquement. Les lignes de code à côté du graphe permettent de reproduire ce graphique. On utilise le package ggplot2 pour construire le graphique, et donc la fonction ggplot applique sur le jeu de données ozone, ensuite on précise grâce à la fonction aes la variable qui sera sur l'axe des x, ici T9, et celle sur l'axe des ordonnées, ici le maximum d'ozone. geom_point permet de dire qu'on veut dessiner les points, et geom_smooth qu'on veut dessiner une courbe, et ici comme method= 'lm' on dessine une droite (lm veut dire linear model)n et se = FALSE indique qu'on ne dessine pas la zone de confiance. Et enfin ggtitle permet de préciser le titre du graphe. Il est possible d'écrire un modèle de régression simple entre le maximum d'ozone et la température à 9h. Plusieurs écritures sont possibles. Mais en toute rigueur on écrit Y_i = beta_0+beta_1x_i+epsilon_i. L'indice i correspond au numéro de l'individu statistique. Dans notre exemple, Y_i sera le maximum d'ozone le jour i, et on écrit que cette valeur est égale à une constante beta_0, plus beta_1 fois la température à 9h ce même jour i et plus un résidu epsilon i. On considère alors que la relation est la même pour tous les individus avec les mêmes beta_0 et beta_1. Des hypothèses sont faites sur les résidus, et ce sont les hypothèses classiques en stat, à savoir que les résidus ont une espérance nulle, ils ont tous la même variance siagma² et ils sont indépendants 2 à 2. Pour généraliser ce modèle qui n'a qu'une seule variable explicative en un modèle de régression avec plusieurs variables explicatives, on va simplement étendre l'écriture en ajoutant les effets de variables w, z etc. Chaque variable est multipliée par un nouveau coefficient beta. On aura donc beta_0 + beta_1 x_i + beta_2 w_i + … + beta_p z_i et on ajoute toujours à la fin un résidu epsilon_i. En effet, même si on ajoute beaucoup de variables, il reste un résidu epsilon_i, c'est-à-dire une part de variabilité que l'on n'explique pas. Si le nombre de variables est grand, on va plutôt écrire les variables x en les numérotant x_i1 sera la ième de la valeur de la variable explicative x_1, x_i2 la ième valeur de la valeur de la variable explicative x_2, etc, jusqu'à x_ip la ième valeur de la valeur de la variable explicative x_p. Pour les résidus epsilon_i, les hypothèses sont à nouveau les hypothèses usuelles à savoir qu'ils ont une espérance nulle, la même variance sigma² et sont indépendants 2 à 2. Voilà comment les effets des variables explicatives quantitatives seront pris en compte dans le modèle.

Diapositive 6 :

Si on s'intéresse maintenant à la liaison entre une variable quantitative, toujours le maximum d'ozone, et cette fois une variable qualitative comme la direction du vent, on représentera les données avec une boîte à moustaches pour chaque modalité de la variable qualitative. Ce graphe permet à nouveau de repérer d'éventuelles données aberrantes, et donne une idée des différences de distribution qu'il peut y avoir pour la variable réponse lorsque les données proviennent de différentes modalités pour la variable qualitative. Là encore, on utilise le package ggplot2 et la fonction

ggplot que l'on applique sur notre jeu de données. On précise ensuite la variable x qualitative, ici le vent, la variable y quantitative ici le maximum d'ozone et on précise avec fill qu'on colorie les boîtes en fonction de la variable vent, et avec col on précise qu'on colorie les points en fonction de la variable vent également. geom_boxplot indique que l'on dessine des boîtes à moustaches, en supprimant les points aberrants avec outlier.spage=NA ; on supprime les points aberrants, ceux qui sont en dehors de la boîte à moustaches, car on dessine ensuite tous les points, y compris ceux aberrants. Et on colorie la boîte avec une certaine transparence en mettant alpha=0.4 ; il faut mettre pour la transparence une valeur entre 0 et 1 pour avoir plus ou moins de transparence. Plutôt que geom_point, on utilise ici geom_jitter qui permet de bouger légèrement les abscisses des points ce qui évite que tous les points correspondant à une même direction du vent aient la même abscisse et soient plus ou moins superposés. Enfin, un titre est mis au graphique.

Comment s'écrit le modèle ? Simplement en écrivant que le maximum d'ozone dépend de la direction du vent. Et plus précisément que le maximum d'ozone du jour j pour une direction du vent i, c'est un maximum d'ozone moyen mu, plus un effet alpha_1 si le vent vient de l'est, alpha_2 s'il vient du nord, alpha_3 s'il vient de l'ouest et alpha_4 s'il vient du sud. Donc selon la direction du vent, on ajoute la valeur spécifique qui convient. Et enfin, on a toujours un résidu, donc quelque chose que l'on n'explique pas. Mathématiquement, on écrira que la réponse Y pour le $j^{ème}$ individu qui prend la modalité i pour la variable qualitative est égal à une valeur moyenne mu, plus une valeur alpha_i qui dépend donc de la modalité de la variable qualitative (par exemple, si le vent vient de l'est le maximum d'ozone sera de +4 par rapport au maximum d'ozone moyen, ou -2 si le vent vient u nord, etc.). Et nous avons ensuite le résidu epsilon_ij qui correspond à l'écart entre notre réponse Y_ij et ce que prévoit le modèle, à savoir mu+alpha_i. Et sur ces résidus, on fait à nouveau les hypothèses usuelles : résidus d'espérance nulle, de même variance $sigma^2$, et des résidus indépendant 2 à 2. Si maintenant on étend ce modèle d'analyse de variance à 1 facteur à un modèle d'analyse de variance à 2 facteurs, il suffit d'ajouter un effet beta_k pour le $2^{ème}$ facteur. On utilise alors un indice k supplémentaire et on écrit que la réponse Y pour le $k^{ème}$ individu qui prend à la fois la modalité i pour le 1^{er} facteur et la modalité j pour le $2^{ème}$ facteur est égale à un effet moyen mu, plus un effet alpha_i qui dépend de la modalité i que prend l'individu pour le 1^{er} facteur, et un effet beta_j qui dépend de la modalité j que prend l'individu pour le $2^{ème}$ facteur. Et on a toujours un résidu epsilon ijk. Ici par exemple, pour le $k^{ème}$ jour de vent d'est où il pleut, on aura un effet moyen, plus un effet alpha_est plus un effet beta_pluie. Et à nouveau les hypothèses classiques pour les résidus. Et il n'y a aucune difficulté à avoir plus de variables explicatives. On va simplement ajouter les effets de chaque variable explicative. Nous avons vu pour l'instant les effets de variables quantitatives sur la réponse, ou les effets de variables qualitatives sur la réponse. Cependant, nous avons pour l'instant considérer que les effets de chaque variable ne font que s'additionner. Voyons maintenant des effets conjoints de deux variables sur la réponse.

Diapositive 7 :

Et commençons par voir l'effet conjoint d'une variable quantitative et d'une variable qualitative sur la variable réponse. On va construire un graphique avec la variable réponse en ordonnée, et en abscisse on met la variable quantitative, ici la température à 9h. On va alors simplement colorier les points en fonction de la variable qualitative, ici la direction du vent. Et on peut dessiner une droite de régression par modalité de la variable vent, i.e. une droite de régression par sous-nuage de points. L'intérêt est alors de voir si l'effet de la température à 9h sur le maximum d'ozone est le même pour toutes les directions du vent ou bien si l'effet de la température à 9h sur le max d'ozone dépend de la direction du vent. Si les pentes des quatre droites étaient identiques, donc si les droites étaient parallèles, alors cela signifierait que l'effet de la température à 9h sur le maximum d'ozone est le même quelle que soit la direction du vent. Autrement dit, un degré de plus à 9h conduit à la même augmentation ou diminution du maximum d'ozone, et cela quelle que soit la direction du vent. Ici les droites ne sont pas parfaitement parallèles, et on a l'impression que l'effet de la température à 9h est moins important quand le vent vient de l'est par rapport aux 3 autres directions du vent. L'augmentation du maximum d'ozone est moindre pour une augmentation de la température de 1 degré à 9h puisque la pente est plus faible. Cependant, ce graphique permet de visualiser l'effet conjoint de la température et de la direction sur le maximum d'ozone, mais nous aurons besoin d'un test statistique pour décider si l'effet conjoint est significatif ou non. On parle d'interaction de la température et du vent, et on testera si l'effet de l'interaction est significatif ou non. Pour construire le graphique, on précise qu'en x on met la variable de vent, en y le maximum d'ozone, on colorie les individus avec une couleur différente pour chaque direction du vent, et on groupe les individus

par direction du vent ce qui permet ensuite de faire des régressions par direction du vent. geom_smooth permet de dessiner la droite de régression, la droite de régression car on met method= « lm » - lm pour linear model. On met se=FALSE pour dire qu'on ne veut pas dessiner de zone de confiance.

L'écriture du modèle donne la variable réponse (le maximum d'ozone) et précise que ce maximum d'ozone dépend de la direction du vent, de la température à 9h, et de l'effet conjoint du vent et de la température à 9h. Cela signifie qu'on n'a pas seulement l'effet du vent, et l'effet de la température, mais bien un effet conjoint en plus. Plus précisément, le maximum d'ozone s'écrit comme un effet moyen, plus un effet lié à la direction du vent i.e. une constante spécifique selon la direction du vent (alpha_1, alpha_2, alpha 3 ou alpha_4) ensuite on ajoute un effet linéaire de la température à 9h en précisant qu'il y a un effet moyen (beta) de la température à 9h plus un effet spécifique (gamma_i) qui dépend de la direction du vent. Autrement dit, la pente de la droite de régression est égale à beta en moyenne, mais selon la direction du vent on ajoute à cette pente une valeur gamma_i, qui est positive ou négative. Ainsi la pente de la droite de régression est plus ou moins forte et donc, l'augmentation de 1 degré de la température à 9h sur le maximum d'ozone sera plus ou moins important selon la direction du vent.

Mathématiquement, le modèle s'écrit Y_ij la réponse pour le $j^{ème}$ individu de la modalité i pour la variable qualitative, est égal à un effet moyen mu, un effet alpha_i lié à la direction du vent pour notre exemple, puis une pente moyenne beta plus un écart de pente gamma_i positif ou négatif selon que la pente est plus ou moins forte que la moyenne, et cette pente est multipliée par x_ij, c'est-à-dire la valeur du $j^{ème}$ individu qui prend la modalité i. Pour nous ce sera la température à 9h du $j^{ème}$ jour quand la direction du vent est i. Et on a en plus les résidus avec les hypothèses usuelles pour les résidus.

Diapositive 8 :

Visualisons maintenant l'effet conjoint de deux variables qualitatives sur la variable réponse. On va parler de l'effet de l'interaction de 2 variables qualitatives sur la variable réponse. Notez bien que, dans le dictionnaire, la définition courante de l'interaction est « la réaction réciproque de 2 phénomènes l'un sur l'autre » : deux phénomènes, ou variables, seulement entrent en jeu et ont chacun un impact sur l'autre. En statistique, l'interaction met en jeu 3 variables et on s'intéresse à l'effet conjoint de 2 variables sur une $3^{ème}$, la variable réponse.

On va construire un graphique avec la variable réponse en ordonnée, en abscisse une des variables qualitatives, et on construit une ligne brisée pour chaque modalité de l'autre variable qualitative. Le choix de la variable à mettre en abscisse et de la variable utilisée pour construire les lignes brisées peut être intervertit. On visualise alors de façon un peu différente la même interaction, et on peut donc faire les 2 graphiques et choisir celui qui est le plus facile à interpréter. Souvent, on préférera mettre en abscisse la variable qualitative qui a le plus de modalités pour avoir moins de lignes brisées. Comment sont construits les lignes brisées ? On va d'abord positionner les points : un point a pour ordonnée la moyenne des valeurs de y pour une combinaison des 2 variables qualitatives. Par exemple pour le couple vent d'est avec pluie, on a ce point, et pour le couple vent d'est sec, on aura ce point. On positionne alors autant de points qu'il y a de combinaisons de modalités des 2 variables qualitatives puis on relie les points d'une même de la variable qualitative pour constituer une ligne brisée. Si les lignes brisées sont parallèles, alors cela signifie qu'il n'y a pas d'effet de l'interaction sur la variable réponse. Et au contraire, s'il y a des écarts au parallélisme, cela signifie qu'une interaction est présente. Par exemple ici, l'écart de max d'ozone entre avec et sans pluie est beaucoup plus important quand le vent vient de l'est, par rapport aux écarts quand le vent vient d'une autre direction. Cela signifie que lorsque le vent vient de l'est, l'effet de la pluie ou du temps sec est exacerbé et le maximum d'ozone est particulièrement élevé quand le temps est sec ou particulièrement faible quand le temps est pluvieux.

Pour construire ce graphique avec ggplot2, on utilise le package dplyr qui permet de grouper les données par combinaison de vent et pluie avec la fonction group_by. Ensuite on calcule la moyenne du max d'ozone pour chaque combinaison avec la fonction summarize. Puis, pour construire le graphique, on place en abscisse le vent et en ordonnée le max d'ozone, on colorie en fonction de la pluie et on groupe en fonction de la pluie ce qui permettra d'avoir une ligne brisée par modalité de la variable pluie. On utilise ensuite geom_line pour tracer les lignes et geom_point pour mettre les points. On ajoute ensuite un titre et les libelles des axes.

Pour écrire le modèle, on dit simplement que le maximum d'ozone est expliqué en fonction du vent, de la pluie et de l'interaction vent-pluie symbolisée par les « : ». Plus dans le détail, on écrit que le maximum d'ozone est égal à un effet moyen mu, plus un effet alpha_i selon la direction du vent. Donc si le vent vient de l'est, on ajoute alpha_1, s'il vient du nord alpha_2, etc. C'est donc l'ajout d'une constante spécifique selon la direction du vent. Même chose ensuite avec la pluie, on ajoute beta_1 s'il pleut et beta_2 s'il fait sec. Enfin, pour l'interaction, on ajoute une valeur spécifique pour le couple de modalité vent-pluie qui est présent. Si le vent vient de l'est et qu'il pleut, on ajoute alphabeta_11. Et on aura une valeur différente alphabeta_ij pour chaque couple. Et bien entendu, il reste un résidu.

Plus formellement, le modèle s'écrit Y_ijk la réponse pour le $k^{ème}$ individu qui prend la modalité i pour la $1^{ère}$ variable qualitative et la modalité j de la $2^{ème}$ variable qualitative, est égal à un effet moyen mu, un effet alpha_i lié à la direction du vent pour notre exemple, un effet beta_j lié à la pluviométrie, puis un effet alphabeta_ij lié au couple direction du vent i et pluviométrie j. Et on a en plus les résidus avec les hypothèses usuelles pour les résidus.

Diapositive 9 :

Nous avons vu 4 types d'effets possibles sur une variable réponse. Ces 4 types d'effet sont les plus couramment utilisés et permettent de construire beaucoup de modèles. Nous retrouvons donc l'effet linéaire d'une variable quantitative sur la variable réponse, l'effet d'une variable qualitative à plusieurs modalités, l'effet joint d'une variable quantitative et d'une variable qualitative sur la variable réponse, effet joint qui permet de différencier l'effet linéaire de la variable quantitative sur la réponse selon les modalités de la variable qualitative. Et enfin, l'effet de l'interaction de 2 variables qualitatives sur la variable réponse. Nous pouvons alors construire des modèles plus ou moins complexes en combinant ces 4 types d'effet selon bien sûr la nature des variables explicatives, mais aussi les effets que l'on souhaite inclure dans le modèle. On peut ne pas vouloir mettre certaines interactions dans le modèle si on sait a priori qu'il n'y a pas d'effet de ces interactions. Et si on n'a pas de connaissance a priori, on mettra l'interaction dans le modèle et on prendra une décision sur l'effet de l'interaction grâce au test statistique qui sera mis en place.

Ces quatre types d'effets et leur interprétation sont très importants à comprendre car toute la difficulté réside dans l'interprétation des résultats, les calculs étant faits par l'ordinateur assez simplement. Donc bien comprendre ce que représente chacun de ces effets et comment le visualiser et l'interpréter est crucial.

Diapositive 10 :

Voici un tableau récapitulatif avec les noms des modèles selon la nature des variables explicatives. Dans tout ce cours, la variable réponse est quantitative. Quand il y a une une variable explicative et que celle-ci est quantitative, il s'agit d'une régression linéaire simple : si la variable explicative est qualitative, il s'agit d'une analyse de variance à 1 facteur. Dans le cas particulier où il n'y a que deux modalités, l'analyse de variance à 1 facteur équivaut à faire une comparaison de moyennes (avec variances égales). Si maintenant il y a plusieurs variables explicatives et que toutes sont quantitatives, on parle de régression linéaire multiple ; si toutes les variables explicatives sont qualitatives il s'agit d'une analyse de variance à K facteurs. Enfin si certaines variables explicatives sont quantitatives et d'autres sont qualitatives, on parle d'analyse de covariance. A chaque fois la variable réponse est quantitative et tous ces modèles sont des modèles linéaires. Si la variable réponse est qualitative, le plus souvent binaire (malade versus non-malade), on peut retrouver les mêmes 4 effets pour les variables explicatives selon la nature des variables et la présence ou non d'interactions. Mais la modélisation est différente et on parlera de régression logistique, méthode qui ne fait pas l'objet de ce cours. La régression logistique s'intéresse à modéliser la probabilité de survenue (ou non) d'un évènement, ou plus précisément modélise le logarithme du ratio de la probabilité de survenue sur la probabilité de non-survenue d'un évènement.

Diapositive 11 :

Reprenons l'écriture du modèle de régression linéaire multiple. En rouge, apparaissent les paramètres du modèle qui doivent être estimés à partir des données. Le nombre de paramètres à estimer est égal au nombre de variables explicatives + 1 car il faut estimer la constante.

Pour estimer tous ces paramètres, nous réécrivons le modèle sous une forme matricielle Y=Xbeta+E, avec Y le vecteur des valeurs observées, X la matrice qui contient toutes les valeurs des variables explicatives (plus précisément, une

1ère colonne avec uniquement des 1 qui permettra d'estimer la constante beta_0, puis une colonne pour chaque variable explicative) ; beta est le vecteur qui regroupe tous les coefficients du modèle à estimer, et E est le vecteur des résidus. En plus de l'équation Y=Xbeta+E, on retrouve les hypothèses sur les résidus, à savoir qu'ils sont sans biais (ce qui est donné par l'espérance de E qui vaut 0), de même variance sigma², et indépendants 2 à 2, ce qui est donné par la variance de E, qui est une matrice de variance covariance qui a sur la diagonale les variances et hors diagonale les covariances entre 2 résidus. Comme on a sigma² fois la matrice identité, on retrouve bien sur la diagonale que tous les résidus ont la même variance sigma² et hors diagonale on retrouve que les résidus sont indépendants 2 à 2. Si on lit le modèle écrit sous forme matricielle, on lit pour la 1ère ligne que Y_1 est égal à 1*beta_0 + x_11 * beta_1 + … + x_1p * beta_p + epsilon 1. En lisant cela, on retrouve l'équation donnée par l'écriture du modèle sous forme indicé pour i=1.

C'est un peu plus difficile à voir, mais les modèles d'analyse de variance et d'analyse de covariance peuvent aussi s'écrire sous la forme Y=Xbeta+E. Et c'est parce que c'est la même écriture qu'on parle de modèles linéaires quelle que soit la nature des variables explicatives et que … les fonctions des logiciels permettant d'estimer les coefficients des modèles sont les mêmes. Pour les variables qualitatives, le codage nécessite d'utiliser plusieurs colonnes pour une même variable, plus précisément autant de colonnes que de modalités. Et pour les interactions entre variables qualitatives, on aura besoin d'autant de colonnes que de couples de modalités. Nous verrons plus tard dans ce cours que des contraintes seront nécessaires pour bien estimer tous les paramètres et qu'il est possible d'utiliser I-1 colonnes pour une variable à I modalités, et (I-1)(J-1) colonnes pour les interactions de 2 variables à I et J modalités.

Diapositive 12 :

Pour estimer les paramètres, nous utilisons le critère des moindres carrés qui consiste à minimiser la somme des carrés des écarts entre les valeurs observées pour la variable Y et les prévisions du modèle.

Pour donner l'idée, on cherche à estimer le vecteur de paramètres beta tel que Xbeta soit le plus proche possible de Y. Si on multiplie à gauche par la transposée de X, on a X'Xbeta est le plus proche possible de X'Y. Et si X'X est inversible alors on peut dire qu'on estime beta par beta chapeau = (X'X)-1 X'Y.

Cet estimateur est sans biais, et la variance de cet estimateur est connue et dépend de X'X, et donc simplement de la matrice X c'est-à-dire du choix des expériences que l'on effectue. Ceci est très important, notamment si on peut planifier les expériences. Dans ce cas, on choisira des expériences telles que X'X prenne des valeurs les plus faibles possibles. La variance de beta chapeau dépend aussi de sigma² et donc de la variance des résidus, ie.e la variance de ce que le modèle n'explique pas.

Comment estimer cette variance des résidus justement ? En calculant la somme des carrés des résidus, et en divisant par les degrés de liberté associé. Les ddl se calculent toujours de la façon suivante : c'est le nombre de données moins le nombre de paramètres estimés à partir des données. Donc pour la régression, ce sera n-(p+1) car n est le nb de données, et p+1 est le nb de paramètres que l'on estime.

Diapositive 13 :

Une fois le modèle construit, une question légitime est de se demander à quel point le modèle explique la variable que l'on cherche à étudier. Les valeurs des Y observées varient, et cette variabilité des valeurs de Y peut se décomposer en une somme de la variabilité expliquée par le modèle, et de la variabilité non-expliquée par le modèle, i.e. la variabilité résiduelle. Ainsi, la variabilité totale de Y, qui est égale à la variance de Y multiplié par (n-1) est bien expliquée par le modèle si la variabilité résiduelle est petite et donc si les prédictions du modèle sont proches des valeurs observées. On calcule alors le R^2, appelé le coefficient de détermination, qui est la part de la variabilité totale de la réponse Y qui est expliquée par le modèle. Ce R^2 varie entre 0, le modèle n'est pas bon du tout, à 1 le modèle prévoit parfaitement les valeurs observées. Si la part de variabilité expliquée par le modèle est grande, le modèle s'ajuste bien aux données observées. Attention toutefois : un modèle peut bien s'ajuster à des données, i.e. « coller aux données observées », mais ne pas être un bon modèle pour prédire de nouvelles données. Cela arrive notamment quand le modèle contient beaucoup de variables et nécessite donc l'estimation de nombreux paramètres. Plus le nombre de paramètres estimés augmente, plus il est facile de « s'ajuster aux valeurs observées », mais pour autant le

modèle n'est pas nécessairement bon pour prédire de nouvelles valeurs. Les modèles qui prédisent bien de nouvelles valeurs sont des modèles qui s'ajustent bien et qui ont nécessité l'estimation d'un nombre raisonnable de paramètres.

Cette variabilité expliquée par le modèle peut être décomposée en parts de variabilité expliquée par chacune des variables explicatives et interactions. Cependant, il y a une difficulté ici car deux variables explicatives, ou plus de 2 d'ailleurs, peuvent expliquer à peu près la même chose, et donc à peu près la même part de variabilité de Y. C'est le cas si 2 variables quantitatives sont très corrélées par exemple. A quelle variable attribuer alors cette variabilité sachant que lorsque le modèle contient les 2 variables la variabilité n'est expliquée qu'une seule fois ? Il y a alors 2 façons de décomposer la variabilité expliquée par le modèle. La 1ère façon est de calculer la part de variabilité expliquée par la 1ère variable du modèle, puis, la part de la variabilité expliquée par la 2ème variable et qui n'a pas encore été expliquée par la 1ère, puis la part expliquée par la 3ème variable non encore expliquée par les 2 premières, etc. L'avantage de cette décomposition est que la variabilité expliquée par l'ensemble du modèle est égale à la somme des variabilités de chaque variable. L'inconvénient majeur est que l'ordre d'apparition des variables dans le modèle modifie la part de variabilité expliquée par une variable. Une autre façon de décomposer la variabilité du modèle est de calculer la part de variabilité expliquée exclusivement par une variable. Ainsi, l'ordre d'apparition des variables dans le modèle ne modifie pas la variabilité expliquée par la variable. En revanche, si une part de variabilité est expliquée par plusieurs variables, alors cette variabilité commune n'apparaît pas dans la décomposition. La somme des variabilités de toutes les variables sera inférieure à la variabilité du modèle. Une variabilité commune est expliquée par plusieurs variables quand les variables explicatives ne sont pas indépendantes.

Quand les données sont équilibrées, i.e. quand les variables explicatives sont indépendantes 2 à 2, la variabilité du modèle se décompose parfaitement et la somme des variabilités expliquées par chaque variable est égale à la variabilité expliquée par le modèle. L'interprétation des résultats est alors beaucoup plus facile.

Diapositive 14 :

Pour construire le modèle et obtenir la variabilité expliquée par chaque variable, on utilise la fonction LinearModel du package FactoMineR. On écrit alors la formule du modèle avec la variable à expliquer (le maximum d'ozone), puis « ~ », et enfin les variables explicatives séparées par un « + », donc on écrit par exemple maxO3~T9 + T12 +T15 + … + vent + pluie en mettant toutes les variables quantitatives et qualitatives.

Quand le nombre de variables explicatives est grand, on peut écrire « ~. » car cela indique que l'on utilise toutes les variables du jeu de données comme variables explicatives (sauf bien entendu le maximum d'ozone qui est la variable réponse). Les sorties du modèle donnent d'abord l'écart-type résiduel qui vaut 14.51, et a 97 ddl. Le R^2 du modèle vaut 0.7686 ce qui signifie que 76.86% de la variabilité du maximum d'ozone est expliquée par ce modèle. Ensuite, un test est construit pour tester la significativité du R^2. Nous reviendrons sur ce test. Et ensuite, nous avons un tableau avec en ligne chaque variable, et dans la colonne SS (pour sum of squares) les sommes de carrés, ce qui correspond à la variabilité expliquée exclusivement par chaque variable. Là encore, nous avons des tests que nous allons décrire ultérieurement.

Diapositive 15 :

Le modèle ainsi construit contient toutes les variables que nous avons choisies comme variables explicatives et toutes les interactions que nous avons pu considérer. On peut alors se demander si tous ces effets sont utiles ou si certains sont superflus. On peut aussi se demander si aucune variable n'est utile, auquel cas les variables explicatives du modèle seraient très mal choisies. Pour cela, des tests existent et ils sont tous construits selon le même principe. La question que l'on se pose est alors « est-ce que cet ensemble de V variables apporte des informations complémentaires intéressantes sachant que les autres variables sont dans le modèle ? » Notez bien ici que l'on teste l'intérêt d'un ensemble de variables sachant que d'autres variables sont déjà présentes dans le modèle, ce qui revient à tester si les variables de l'ensemble V apporte un plus. Mathématiquement, on va tester une hypothèse H0 contre une hypothèse alternative H1. L'hypothèse H0 est : tous les coefficients associés aux variables de l'ensemble V sont nuls. Et l'hypothèse alternative est que l'un au moins des coefficients n'est pas nul. La statistique de test utilisée pour ce test est le ratio entre le carré moyen de l'ensemble V et le carré moyen de la résiduelle. Le carré moyen d'un effet est simplement la variabilité expliquée par cet effet divisé par les ddl de cet effet. Dans cette statistique de test, le carré

moyen résiduel sert de comparaison. Si la variabilité expliquée par l'ensemble V de variables est du même ordre de grandeur que la variabilité résiduelle, alors cela veut dire que l'ensemble V de variables explique peu de chose (à peu près autant que ce que la résiduelle explique), c'est-à-dire autant que la variabilité que l'on n'arrive pas à expliquer. Dans ce cas, on aura envie de dire que l'ensemble V de variables n'apporte pas une information significativement intéressante pour mieux expliquer Y. Et au contraire, si l'ensemble V apporte beaucoup plus d'information que la résiduelle, i.e. si le Fobs est grand, alors on aura envie de dire que l'ensemble V apporte une information significativement intéressante pour comprendre Y. La loi de la statistique de test sous l'hypothèse H0, est une loi de Fisher avec les ddl du numérateur et les ddl du dénominateur.

Avec ce test, retenir l'hypothèse H0 revient à considérer que le sous-modèle (sans les variables de l'ensemble V) est aussi bon que le modèle complet. Le test conduit à choisir le sous-modèle si la variabilité supplémentaire expliquée par l'ensemble V de variables est du même ordre de grandeur que la variabilité résiduelle (i.e. la stat F de Fisher n'est pas significativement différente de 1). Le plus souvent, l'ensemble V est restreint à 1 seule variable, ce qui revient à tester si cette variable apporte une information supplémentaire intéressante sachant que les autres variables sont déjà présentes dans le modèle. L'autre situation classique est au contraire d'avoir l'ensemble V qui contient toutes les variables. L'hypothèse H0 correspond au modèle nul, c'est-à-dire au modèle sans variable explicative. Ce modèle nul n'est absolument pas intéressant pour prédire et indique, s'il est préféré au modèle complet, que le modèle complet n'apporte pas d'information complémentaires intéressantes par rapport à un modèle qui n'explique rien. Et par conséquent, le modèle complet n'est pas intéressant et donc les variables explicatives sont mal choisies.

Si l'ensemble V contient toutes les variables, le test revient à tester si le R^2 est significativement différent de 0, i.e. si toutes les variables explicatives ont un coefficient égal à 0 et sont donc inutiles, ou au contraire si au moins une variable est utile pour prédire Y.

Pour construire le test, on calcule la variabilité expliquée par l'ensemble V de variables sachant que toutes les autres variables sont déjà dans le modèle. On est donc bien en train de voir si l'ensemble V de variables apporte un plus par rapport aux autres variables présentes dans le modèle. Et on divise cette variabilité par les degrés de liberté associés à cet ensemble de variables. Pour calculer les ddl il suffit de sommer les ddl de chaque variable et interactions présentes dans l'ensemble V sachant qu'une variable quantitative à 1 ddl, une variable qualitative à I-1 ddl et une interaction a comme ddl le produit des ddl de chaque facteur.

Il peut être intéressant comme exercice d'écrire le test dans les cas particulier suivant : tester l'effet d'une variable qualitative, le test d'une interaction, le test de toutes les variables.

Diapositive 16 :

Pour l'instant, nous avons construit un modèle utilisant toutes les variables explicatives qui nous semblent pour influer sur Y, et nous avons également ajouter des interactions qui peuvent également avoir un effet sur Y. Cependant, ce modèle peut nécessiter l'estimation de nombreux paramètres et il est essentiel de sélectionner les variables « utiles » qui seront conservées dans le modèle. Cela permet d'une part de simplifier la compréhension du phénomène que l'on étudie en supprimant quelques variables, mais cela permet aussi de construire un modèle plus parcimonieux, pour lequel on aura estimé moins de paramètres, et qui plus performant d'un point de vue prédictif. Se pose alors la question de comment sélectionner un « bon » modèle ? Et qu'est-ce qu'un « bon » modèle ? Un bon modèle est un modèle qui permet de bien comprendre les relations entre les variables explicatives et la variable réponse, et qui permet également de bien prédire la variable réponse pour de nouvelles observations.

Pour sélectionner un « bon » modèle, on peut alors utiliser 2 grandes stratégies. La première est de sélectionner le sous-modèle pour lequel la probabilité critique associée au test de significativité du R^2 est la plus petite. On choisit alors le sous-modèle pour lequel on rejette le plus fortement l'hypothèse que le sous-modèle n'est pas intéressant … ce qui indique que le modèle est intéressant. Une autre façon de faire est de sélectionner le modèle à partir d'un critère comme le critère d'Akaïke, noté AIC, ou encore le critère d'information bayésien, noté BIC pour bayesian information criterion. Le sous-modèle qui a l'AIC ou le BIC le plus petit est le meilleur. Sélectionner selon l'AIC et le BIC peut conduire à des sous-modèles différents, le critère BIC retenant des sous-modèles avec moins de variables. On utilisera de préférence le critère AIC quand on veut bien prédire, et le critère BIC quand on veut sélectionner les

variables explicatives et comprendre le phénomène. Ces 2 critères fonctionnent selon le même principe qui est de retenir un modèle avec la variabilité résiduelle minimum (en toute rigueur l'opposé du logarithme de la vraisemblance minimum) mais tout en ayant peu de paramètres (les critères sont d'autant plus pénalisés que le nombre de variables du modèle est grand). Quand les variables explicatives sont quantitatives et nombreuses, on recommande d'utiliser le critère BIC, et quand des variables explicatives sont qualitatives, on recommande plutôt l'AIC.

Comment maintenant trouver le sous-modèle qui a la p-value la plus petite pour le test du R^2, ou qui a l'AIC le plus petit ou le BIC le plus petit ? Une stratégie très basique consiste à construire tous les sous-modèles possibles, à calculer le critère qui nous intéresse pour chaque sous-modèle, et à retenir le sous-modèle qui minimise notre critère. Cependant, le nombre de sous-modèles possibles devient très vite extrêmement grand et les temps de calcul prohibitifs dès que le nombre de variables explicatives, variables explicatives ou interactions, du modèle est grand. Pour contourner ce problème, plusieurs stratégies sont possibles. La première, appelée méthode descendante, consiste à construire le modèle complet, puis à supprimer la variable (ou l'interaction) la moins intéressante donc qui diminue très peu le critère qui nous intéresse (rappelons-nous que, quel que soit le critère que l'on utilise, on cherche à le minimiser). Et on itère jusqu'à ce que, quelle que soit la variable que l'on supprime, le critère ne diminue plus. On a alors trouvé un sous-modèle qui minimise notre critère parmi tous les sous-modèles qui ont été explorés. Cela ne garantit pas que nous avons trouvé le sous-modèle avec le critère minimal, cela reste un bon sous-modèle. Une autre stratégie consiste à faire l'inverse : partir du modèle nul sans variable, et ajouter la variable qui améliore (i.e. diminue) le plus possible notre critère d'intérêt. Ensuite on garde ce sous-modèle, et on itère en ajoutant la variable (ou l'interaction) qui améliore le plus le critère. Et on continue jusqu'à ce que, quelle que soit la variable ou l'effet ajouté, le critère se met à augmenter. Cette méthode est appelée méthode ascendante. On peut ensuite avoir une méthode nommée stepwise, qui combine la méthode descendante et la méthode ascendante, ce qui permet de rajouter une variable qui aurait été supprimée lors des premières étapes, ou à enlever une variable qui aurait été rajoutée et qui n'est plus intéressante une fois que d'autres ont été ajoutées. Cette méthode stepwise est rapide et permet souvent de trouver un sous-modèle proche du meilleur modèle qu'on ne peut obtenir avec certitude qu'avec la méthode de construction exhaustive de tous les sous-modèles.

Diapositive 17 :

Reprenons notre exemple sur les données ozone, et utilisons la fonction LinearModel du package FactoMineR pour choisir un sous-modèle avec le critère BIC. La fonction LinearModel retourne les résultats pour le modèle complet et pour le sous-modèle sélectionné. On voit par exemple que le R^2 du sous-modèle est légèrement plus petit (0.7622 contre 0.7686 pour le modèle complet). Ainsi le pourcentage de variabilité du maximum d'ozone expliqué par le sous-modèle est quasiment le même qu'avec le modèle complet alors qu'on utilise beaucoup moins de variables et donc qu'on estime moins de paramètres. Les valeurs de la probabilité critique, de l'AIC et du BIC du sous-modèle sélectionné sont plus petites que celles du modèle complet. Le sous-modèle sélectionné est écrit ici et contient seulement 4 variables, T12, Ne9, Vx9 et maxO3v.

Diapositive 18 :

On peut ensuite regarder le tableau qui donne pour chaque variable explicative et interaction, les résultats du test de la significativité de cet effet sachant que les autres variables sont présentes dans le modèle. Dans le tableau des tests F donc, on peut lire que toutes les p-values sont inférieurs à 5%, et donc chacune des quatre variables apporte une information significative supplémentaire utile sachant que les autres variables sont dans le modèle.

Dans certains cas, on peut avoir des résultats avec certaines variables ou interactions qui ont une probabilité critique supérieure à 5%. Le test F nous indiquerait que l'on peut supprimer la variable ou l'interaction qui a la p-value la plus élevée alors que le critère BIC a conservé cette variable dans le sous-modèle. Ces situations peuvent arriver et conduisent à une légère contradiction au niveau de l'interprétation sur l'utilité de garder ou non cette variable dans le modèle. Mais cette variable aura généralement peu d'effet sur la prédiction de nouvelles valeurs et la garder ou non dans le sous-modèle ne modifie pas beaucoup les prédictions.

Nous verrons ensuite comment interpréter le 2ème tableau avec les tests T de Student dans la 2ème partie du cours.

Diapositive 19 :

Quelle démarche adopter en modélisation ? La 1ère étape est de lister toutes les variables qui entrent en jeu pour expliquer ou prédire la variable réponse. Cette étape est essentielle et il ne faut surtout pas oublier une variable qui peut influer sur Y. Prendre une variable en trop n'est pas grave car les tests statistiques permettront d'éliminer cette variable. 2ème étape, il faut visualiser les données et notamment les liaisons avec la variable réponse pour s'assurer qu'il n'y a pas de valeurs aberrantes et pour avoir quelques idées sur les liaisons entre variables. Ensuite, on écrit puis construit le modèle en choisissant à la fois les effets et les interactions qui peuvent expliquer la réponse. On ne mettra pas systématiquement, et sans réfléchir, toutes les interactions possibles dans le modèle. On peut souvent écarter a priori certaines interactions quand on connaît le phénomène que l'on étudie. Et surtout, on évite de mettre des interactions d'ordre supérieures à 2. Ensuite, à partir du modèle complet, on sélectionne le sous-modèle qui minimise l'AIC (ou le BIC ou à la main) en supprimant les interactions et effets non utiles, puis on construit ce sous-modèle.

Ensuite, et c'est le plus important, on interprétera les résultats (les effets significatifs comme les non significatifs), on interprétera les coefficients du modèle en faisant bien attention aux possibles confusions notamment pour les variables quantitatives, et enfin on utilise le sous-modèle pour prédire de nouvelles valeurs si on a un objectif de prévision. Nous verrons comment interpréter et prédire dans une 2ème partie de ce cours.

Diapositive 20 :

Après avoir vu comment construire un modèle et comment sélectionner un « bon » modèle, voyons dans cette 2ème partie du cours comment interpréter les résultats et comment prédire grâce au modèle.

Diapositive 21 :

Pour savoir comment interpréter les résultats d'un modèle, revenons dans un premier temps sur le codage qui a permis d'estimer les paramètres du modèle, et sur les contraintes qui sont utilisées pour les variables qualitatives.

Dans le modèle Y=Xbeta+E, la matrice X est la matrice qui contient toutes les valeurs permettant de décrire précisément chaque expérience, mais aussi chaque effet qui sera utilisé dans le modèle (i.e. les interactions si des interactions sont présentes dans le modèle). La matrice X a alors une ligne par individu statistique (une ligne décrit une expérience). Et comme colonne, la matrice X a les colonnes suivantes : tout d'abord, une colonne avec uniquement des 1 qui sert à calculer la constante du modèle, notée mu ou beta_0. Un ddl est donc utilisé ici pour estimer la constante. Chaque variable quantitative correspond à une colonne de X et utilise donc un ddl pour estimer son paramètre correspondant. Chaque variable qualitative est transformée en autant d'indicatrices qu'elle a de modalités (avec le codage 1 si l'individu prend la modalité et 0 sinon). Cependant, il suffit de connaître I-1 indicatrices pour connaître la dernière. Il suffit donc d'estimer I-1 paramètres à partir des données, et donc d'utiliser I-1 ddl, pour estimer tous les coefficients alpha_i. On pose alors une contrainte sur les alpha_i, et le mieux est de prendre la contrainte que la somme des alpha_i est égale à 0. Chaque interaction est codée par autant d'indicatrices qu'il y a de paires de modalités. Comme pour une variable qualitative, il n'est pas nécessaire d'avoir toutes les indicatrices et avec seulement (I-1)(J-1) indicatrices, et donc avec (I-1)(J-1) ddl, on peut estimer tous les paramètres. Les contraintes sont alors : pour tout i, la somme sur j des alphabeta_ij vaut 0, et pour tout j, la somme sur i des alphabeta_ij vaut 0.

Une remarque fondamentale, et qu'il faut toujours garder en tête au moment où l'on interprète les résultats d'un modèle : le choix de la contrainte impacte FORTEMENT l'interprétation. Avec la somme des alpha_i=0, interpréter la valeur d'un coefficient alpha_i revient à comparer ce qui se passe pour la modalité i par rapport à ce qui se passe en moyenne ; plus précisément, la comparaison se fait pas rapport à la moyenne des moyennes par modalité (un coefficient égal à 0 signifie que les résultats de la modalité ne sont pas différents de l'effet moyen). L'autre contrainte classique possible, est de poser alpha_1=0. Cette contrainte est utilisée par défaut pour certaines fonctions de R (par ex. la fonction lm), et lors de l'interprétation du coefficient alpha_i on comparera ce qui se passe pour la modalité i par rapport à ce qui se passe pour la modalité 1 qui est la référence. Cette contrainte est à proscrire quand il y a des interactions car l'interprétation devient très très compliquée.

Diapositive 22 :

Pour interpréter les résultats du modèle, on va se concentrer sur les résultats obtenus grâce au sous-modèle après avoir fait la sélection de variables. Ainsi, certaines variables ou interactions ont pu être supprimées du modèle. On interprète alors la présence des variables et interactions dans le modèle, mais aussi l'absence des effets et interactions qui n'apparaissent plus dans le modèle.

Il y a 2 situations qui rendent l'interprétation bien différente. La première situation où les données sont équilibrées est idéale. Les données sont équilibrées si tous les effets sont orthogonaux 2 à 2. L'interprétation est alors facile car tous les effets s'additionnent et il n'y a pas de confusion entre les effets. La 2ème situation, quand les données sont déséquilibrées, rend l'interprétation beaucoup plus difficile.

C'est parce que l'interprétation est difficile quand les données sont déséquilibrées que, si on le peut, on va chercher à construire des plans d'expériences et donc à récupérer des données, telles que celles-ci soient équilibrées. Il y a donc un pan de la statistique appelée planification expérimentale qui permet de bien choisir les expériences à réaliser.

Lorsque les variables explicatives sont qualitatives, il est relativement facile d'avoir des données équilibrées ou en tout cas peu déséquilibrées. Pour avoir des données équilibrées, il suffit en effet de tester toutes les combinaisons des facteurs un même nombre de fois. En régression ou analyse de covariance, très souvent les données sont déséquilibrées car elles ont été recueillies sans plan d'expérience. Dans notre exemple sur l'ozone, les températures et les nébulosités ont été recueillies et il y a des corrélations qui existent entre ces variables explicatives, ce qui conduit à des données déséquilibrées.

Diapositive 23 :

Quand les données sont équilibrées et dans le cadre de l'analyse de variance, i.e. quand toutes les variables sont qualitatives, on peut décomposer la variabilité de la réponse Y en plusieurs termes. La variabilité de Y, qui est donc la variance de Y multipliée par (n-1) est égale à la somme de 4 variabilités : la variabilité liée au 1er facteur, donc lié au fait que la moyenne des Y pour chaque modalité i n'est pas la même pour tout i, la variabilité du 2ème facteur, la variabilité liée à l'interaction qui est ce terme, et enfin la variabilité résiduelle.

Et si les données sont équilibrées, on estime très facilement chaque paramètre du modèle : l'effet moyen mu est estimé par la moyenne des données ; l'effet de la modalité i du 1er facteur est simplement l'écart entre la moyenne des valeurs pour la modalité i et la moyenne générale (très intuitif tout ça) ; de même l'effet beta_j s'estime par la différence entre la moyenne des valeurs prises pour la modalité j et la moyenne générale. Pour l'interaction, on peut voir ce terme comme (Yij. – Yi..) – (Y.j.-Y…). Le 1er terme, Yij.-Yi.. revient à calculer l'effet de la modalité j, en se restreignant uniquement aux individus qui prennent la modalité i pour le 1er facteur. Quant au terme Y.j.-Y…, il correspond à l'effet de la modalité j, i.e. à notre beta chapeau j. Cette variabilité de l'interaction revient donc à comparer l'effet spécifique de la modalité j quand le 1er facteur prend la modalité i, par rapport à l'effet moyen de la modalité j. Enfin, le dernier terme correspond au résidu du modèle, et donc à l'écart entre valeur observée et prédite par le modèle.

Cette écriture montre que les effets de chaque facteur, et de l'interaction s'interprètent facilement quand les données sont équilibrées car on compare simplement des moyennes : moyenne des valeurs de Y quand le facteur 1 prend la modalité i moins la moyenne des valeurs de Y pour alpha_i par exemple. Si les données sont déséquilibrées, on n'a plus égalité entre alpha_i chapeau et la différence des moyennes, et de même pour tous les autres coefficients du modèle. Et l'interprétation est plus difficile et on ne peut estimer les paramètres du modèle qu'avec la formule matricielle (X'X)-1X'Y.

Diapositive 24 :

Quand les données sont déséquilibrées, les problèmes d'interprétation commencent ! Et la difficulté d'interprétation augmente avec le déséquilibre dans les données. Si les données sont trop déséquilibrées, il sera impossible de distinguer quel effet ou variable explique la variabilité de Y. On parle de confusion ou d'alias en anglais. Donnons un exemple de confusion. Si par exemple sur les données ozone, tous les jours où la température à 9h est basse le vent vient du nord, et tous les jours où la température est haute le vent vient du sud. Si le maximum d'ozone est plus élevé, comment savoir si c'est parce que les températures sont élevées ou parce que le vent vient du sud ? Il y a une confusion

entre T° élevée et vent du sud, et il sera impossible de savoir ce qui fait varier le maximum d'ozone : est-ce le vent ou la température élevée ? Le modèle retiendra peut-être une seule des 2 variables lors de l'étape de sélection de variables, mais lors de l'interprétation, il faudra avoir en tête que l'effet sur le maximum d'ozone peut être dû aux 2 variables.

Avec des données déséquilibrées, la somme des variabilités expliquées par chacun des effets n'est pas égale à la variabilité totale quand on comptabilise uniquement la variabilité expliquée exclusivement par chaque variable.

Quand les variables explicatives sont très liées, i.e. très corrélées pour les variables quantitatives, l'interprétation est très difficile. En général, la sélection de variables évite de mauvaises interprétations qui sont possible comme un signe du coefficient de régression différent du coefficient de corrélation entre la variable explicative et la réponse. Cependant, l'effet d'une variable explicative peut cacher l'effet d'autres variables non sélectionnées dans le modèle.

Diapositive 25 :

On peut construire le test de significativité d'une variable qualitative. Mais ce test est juste un cas particulier du test général présenté à la dispositive 15. La question est de savoir si une variable a un effet sur la variable réponse. L'hypothèse H0 est que, quelle que soit la modalité du facteur d'intérêt, les valeurs de Y ne sont pas significativement différentes, contre il y a au moins une modalité pour laquelle les individus prennent en moyenne des valeurs significativement différentes. Mathématiquement, on teste donc H0, tous les coefficients alpha_i sont égaux à 0 ; contre H1 il y a au moins une modalité i pour laquelle le coefficient alpha_i est différent de 0 (et donc les valeurs de Y sont en moyenne significativement différentes). La statistique de test est le ratio entre la variabilité expliquée par le facteur que l'on teste, divisée par les ddl de ce facteur, sur la variabilité de la résiduelle, divisée par les ddl de la résiduelle. La variabilité résiduelle sert alors de valeur de comparaison pour voir si la variabilité du facteur d'intérêt est grande ou petite. Si la variabilité du facteur est du même ordre de grandeur que la variabilité de la résiduelle, cela signifie que le facteur explique autant de variabilité que la résiduelle, i.e. autant de variabilité … que ce que l'on n'arrive pas à expliquer. Si au contraire, la variabilité du facteur testé est significativement plus grande que la variabilité résiduelle, alors on va dire qu'il y a un effet significatif du facteur. Si l'hypothèse H0 est vérifiée, cette statistique F_obs suit une loi de Fisher avec les ddl du numérateur et les ddl du dénominateur. Et donc, grâce à la p-value associée qui est donnée par tout bon logiciel, on peut prendre une décision entre les deux hypothèses H0 et H1.

Si on veut tester une interaction, on construit exactement le même type de test. L'hypothèse H0 est que tous les coefficients d'interaction sont nuls, contre l'alternative, il y a au moins un couple i-j tel que le coefficient soit différent de 0. Là encore, on a une statistique de test qui va comparer la variabilité de l'interaction avec la variabilité résiduelle. Et sous l'hypothèse H0, la statistique de test suit une loi de Fisher avec les ddl de l'interaction et les ddl de la résiduelle.

Diapositive 26 :

Si maintenant on rejette l'hypothèse H0 et que l'on considère qu'il y a un effet du facteur A, et donc qu'il y a au moins un coefficient alpha_i qui est différent de 0, alors on a envie de savoir quel(s) coefficient(s) alpha_i est différent de 0, ou plutôt quels sont les coefficients alpha_i qui sont différents de 0 ? On va donc construire autant de tests qu'il y a de modalités pour le facteur A, donc I. Et on va tester si le coefficient alpha_1, par exemple, est différent de 0 ou non, ce qui revient à tester si la modalité 1 donne des résultats égaux ou significativement différents de la moyenne. L'hypothèse H0 est alpha_1=0, et H1 alpha_1<>de 0. La statistique de test Tobs est une statistique de Student qui, sous l'hypothèse H0, suit une loi de Student avec les ddl de la résiduelle. On utilisera la p-value pour prendre la décision sur le test.

Remarquons que pour les variables quantitatives, tester si la variable a un effet linéaire significatif revient à tester si le coefficient associé à la variable est égal à 0, ce qu'il est possible de faire avec ce test.

Diapositive 27 :

Une autre façon d'aller plus dans le détail une fois que l'on a rejeté l'hypothèse H0 du test global d'un effet, donc une fois que l'on a affirmé qu'il y a un effet du facteur A, consiste à comparer les moyennes de Y pour différentes modalités du facteur A. Plus exactement, on va comparer les moyennes ajustées. Les moyennes ajustées utilisent les résultats

du modèle et sont simplement les mu chapeau + alpha_i chapeau. Utiliser les moyennes ajustées permet de s'affranchir de l'effet des autres variables, comme si pour les autres variables on travaillait avec une modalité moyenne. Concrètement, si les données sont équilibrées, cela ne change rien car la moyenne et la moyenne ajustée sont égales. Mais si les données sont déséquilibrées, travailler sur les moyennes ajustées neutralise l'effet des autres variables.

On peut alors comparer les modalités 2 à 2 et construire des tests de comparaison par paire. Comme on construit beaucoup de tests, on va utiliser une correction des tests, comme la correction de Bonferroni. Cela évite de rejeter à tort une hypothèse H0 quand on fait de nombreux tests.

La fonction meansComp de FactoMineR permet de construire ces tests de comparaison. On peut donner un exemple avec des données où la saveur sucrée est évaluée pour 6 chocolats par 21 juges. On construit le modèle avec la fonction LinearModel de FactoMineR et on précise que la variable sucré est expliquée en fonction d'un effet chocolat, d'un effet juge et de l'interaction chocolat:juge. Ensuite, on va comparer les moyennes (ajustées) des chocolats grâce à la fonction meansComp et la correction de Bonferroni. La fonction meansComp retourne d'abord un tableau avec toutes les moyennes ajustées et leur intervalle de confiance. Ensuite, la fonction classe les chocolats par moyenne croissante, et indique avec des lettres quelles sont les moyennes qui sont égales, plus exactement non significativement différentes. Par exemple le chocolat 1 est dans le groupe a, et c'est le seul à être dans le groupe a. Il est donc significativement différent de tous les autres chocolats. Le chocolat 4 est dans le groupe b, il n'est donc pas significativement différent des chocolats 2 et 5 qui sont aussi dans le groupe b. Les chocolats 1, 6 et 3 n'étant pas dans le groupe b, le chocolat 4 est différent de ces chocolats. Quant au chocolat 5, il est dans le groupe b, donc pas significativement différent des chocolats 4 et 2, et il est aussi dans le groupe c, et donc pas significativement différent du chocolat 6. On voit qu'ici certains chocolats peuvent être non significativement de 2 chocolats qui sont eux significativement différent. Ceci s'explique par le fait que le chocolat 5 a une moyenne ajustée de 5.22 qui est entre les moyennes ajustées des chocolats 2 et 6, resp. 4.62 et 5.62. Par contre les chocolats 2 et 6 ont des moyennes ajustées significativement différentes : l'écart entre 4.62 et 5.62 est significatif.

La figure fournie par meansComp résume toutes ces informations. En abscisse, vous avez les moyennes ajustées, en ordonnée le chocolat. Les chocolats sont triés par moyenne ajustée croissante, le point donne la moyenne ajustée et la barre l'intervalle de confiance. A coté de la barre on peut lire les lettres des groupes auxquels le chocolat appartient.

Diapositive 28 :

Un des objectifs de la modélisation est de prédire les valeurs de la variable réponse pour de nouvelles valeurs des variables explicatives. Dans notre exemple, ce serait par exemple de prédire le maximum d'ozone à partir de nouvelles données de température, vitesse de vent et nébulosité. Evidemment, cela nécessite de connaître les températures, vitesse de vent et nébulosité, ce qui n'est pas simple, ou plutôt qui est connu une fois la journée passée … et donc le pic du maximum d'ozone passé. On peut cependant utiliser les données de prévision météo comme entrée, ou encore on peut utiliser le modèle pour prédire la concentration du maximum d'ozone dans des situations d'augmentation des températures par exemple. Pour la prédiction, on va simplement utiliser les coefficients du modèle que l'on a estimés. Bien sûr, on utilise le modèle sans ajouter de résidu car la prédiction ne contient pas de résidu (le résidu serait l'écart justement entre la prévision du modèle et la valeur observée). Donc, dans le listing, on récupère sur la 1ère colonne les estimations des coefficients, et on multiplie par les valeurs des variables explicatives en utilisant simplement l'équation : 12.63 la constante + 2.764 fois la valeur pour la T° à 12h -2.51 fois la nébulosité à 9h + 1.29 fois la vitesse du vent à 9h + 0.35 fois le maximum d'ozone. Ceci est tout à fait intuitif. Sur ordinateur, c'est un peu moins intuitif car il faut saisir les nouvelles données dans un data-frame, puis utiliser la fonction predict sur l'objet résultat de notre model (donc l'objet qui contient les coefficient), ensuite on donne le nom du data frame, ici xnew, qui contient les nouvelles valeurs, et on peut préciser soit qu'on veut un intervalle de confiance de la prédiction avec « pred », soit un intervalle de confiance de la moyenne avec « confidence ». L'intervalle de confiance de la moyenne est d'amplitude plus faible car il donne l'intervalle de confiance d'une moyenne des prédictions pour les valeurs xnew tandis que pred donne l'intervalle pour 1 prédiction, et donc pour un jour donnée. Dans cet intervalle, il y a l'incertitude journalière qui augmente sensiblement l'amplitude de l'intervalle de confiance.

Diapositive 29 :

Un petit commentaire sur les hypothèses qui sont posées et que l'on peut vérifier lorsque l'on construit un modèle. Notez bien que ces hypothèses concernent les résidus du modèle … et donc qu'il faut avoir construit le modèle et calculé les résidus pour savoir si on a le droit de construire le modèle : c'est un peu paradoxal de construire le modèle d'abord pour savoir si on a le droit de le construire ! La 1ère hypothèse est l'égalité de la variance des résidus. On parle de l'homoscédasticité des résidus. Cette hypothèse est importante, et il ne faut pas que la variance des résidus soit grande pour certaines modalités d'une variable qualitative et petite pour d'autres. Ou alors que la variance des résidus soit petite pour les résidus correspondant à certaines valeurs d'une variable quantitative (grandes ou petites) et que la variance des résidus soit grande pour les autres. On peut tester l'égalité de variance grâce à un test de Bartlett.

Une autre hypothèse, beaucoup moins importante à tester, est la normalité des résidus. En fait, cette hypothèse n'est pas importante car il suffit que la distribution des résidus soit à peu près symétrique pour que les tests ne soient pas problématiques. En fait, il faut surtout éviter ici des distributions de résidus très dissymétriques, du type lognormale avec un pic en 0 par exemple. On testera la normalité des résidus grâce à la fonction shapiro .test. J'insiste que c'est sur les résidus que l'on fait le test.

Diapositive 30 :

Revenons plus en détail sur l'exemple ozone. On s'intéresse à comprendre ce qui influence le maximum d'ozone journalier en fonction de données climatiques, et on veut proposer un modèle pour prédire le maximum d'ozone. La variable réponse est donc bien le maximum d'ozone, et les variables explicatives à notre disposition sont la température, la nébulosité et la vitesse du vent à 9h, 12h et 15h ainsi que le maximum d'ozone de la veille. On a de plus la direction du vent et le fait qu'il pleuve ou non. On peut donc construire un modèle avec toutes ces variables, mais doit-on ajouter en plus des interactions ? L'interaction entre la direction du vent et la température à 9h existe si l'effet de la température à 9h sur le maximum d'ozone n'est pas le même selon que le vent vienne de l'est, de l'ouest, du nord ou du sud. Il est difficile de décider a priori si une telle interaction existe. Dans le doute, on choisira de mettre l'interaction dans le modèle, et on utilisera les tests statistiques et la sélection de modèle pour décider. On construit alors un modèle avec tous les effets potentiels possibles, et ici, comme on ne peut pas décider a priori que certaines interactions sont négligeables, ont les met toutes dans le modèle. Pour écrire le modèle, on peut alors donner la formule maxO3 « ~ » et une formule avec tous les effets et interactions. Pour écrire une telle formule, le symbole « * » indique que l'on considère à la fois les 2 effets et l'interaction (A*B indique A+B+A :B). Il faut manipuler le symbole « * » avec parcimonie car il ne faut pas mettre d'interactions d'ordre supérieur à 2 dans le modèle (en toute rigueur, ce serait possible, mais il est très rare que des interactions d'ordre 3 ou plus soient non négligeables et de plus elles sont très difficiles à interpréter). L'écriture de la formule utilise les mêmes règles qu'en mathématique avec les signes « * », les parenthèses ou encore le signe « - ».

Le modèle tel que nous l'avons écrit revient donc à écrire ce modèle avec toutes les interactions, et toutes les interactions avec vent et toutes celles avec pluie. Notez que l'interaction d'un effet avec lui-même revient à écrire simplement l'effet (A :A équivaut à A) ; si vous mettez plusieurs fois le même effet, il ne sera considéré qu'une fois (si on écrit pluie :vent et vent :pluie, l'interaction entre la pluie et le vent ne sera estimée qu'une fois). On voit que le modèle que l'on construit utilise beaucoup d'effets et d'interactions, et par suite nécessite l'estimation de beaucoup de paramètres. Rappelons que les variables quantitatives ont 1 ddl, que la variable vent a 3 ddl (4 directions de vent – 1), et donc toutes les interactions avec vent utiliserons également 3 ddl. La variable pluie ayant 2 modalités, elle n'utilise qu'un ddl, tout comme toutes les interactions de la pluie avec les variables quantitatives. On a donc 11 variables avec 1 ddl, 1 variable à 3 ddl, 10 interactions avec vent qui ont 3 ddl et 10 interactions avec pluie qui ont 1 ddl et l'interaction pluie - vent qui a 3 ddl. On a donc 57ddl + 1 paramètres pour la constante soit 58 paramètres à estimer à partir des données. La sélection de variables et d'interactions diminuera grandement le nombre de paramètres à estimer.

Diapositive 31 :

On l'a vu le nombre de paramètres à estimer est très grand et il est donc important de sélectionner un sous-modèle. En utilisant le critère BIC pour faire la sélection, on retient un sous-modèle beaucoup plus simple avec les variables T9, T15, Ne12, Vx9, maxO3v, vent et les interactions T9-vent et T15-vent. Le R^2 est plus petit avec le sous-modèle, mais ceci est attendu car nous avons estimé beaucoup moins de paramètres.

Diapositive 32 :

Et maintenant, le plus intéressant commence : l'interprétation des résultats. On peut donc voir le tableau des tests globaux de chaque effet, donc de chaque variable explicative interaction. Ce tableau donne la variabilité du maximum d'ozone expliquée exclusivement par chaque effet, mais nous allons surtout commenter la p-value. Les p-value sont inférieures à 0.05, sauf pour la variable vent et les variables T9 et T15. Mais ces variables sont conservées dans le modèle car elles influent sur le maximum d'ozone à travers les interactions T9:vent et T15:vent. On peut donc dire, qu'il y a des effets significatifs de la nébulosité, de la vitesse de vent et du maximum d'O3 de la veille sur le max d'O3. Les interactions significatives permettent de dire qu'il y a des effets de la direction du vent et de la T° sur le maximum d'ozone, mais l'effet de la T° dépend de la direction du vent. Autrement dit, selon la direction du vent, l'effet de la T° est amplifié ou diminué.

On peut aussi commenter les effets non significatifs : la pluviométrie n'est pas un facteur déterminant qui influe sur le max d'03 car elle n'apparaît pas dans le modèle. Une seule nébulosité (Ne12) est conservée dans le modèle : cela ne veut pas dire que les autres nébulosités n'ont pas d'effet. Elles peuvent avoir un effet similaire à celui de la nébulosité à 12h. Donc attention ici lors de l'interprétation à ne pas croire que la nébulosité à 12h influe sur le max d'ozone, et que les deux autres n'influent pas. Il y a de grande chance qu'il y ait une corrélation forte entre toutes les variables de nébulosité, et que le modèle ne retienne qu'une seule variable car l'information portée par les variables de nébulosité est redondante et peut être résumée par une seule variable. C'est la même chose pour la vitesse de vent. Pour la T°, on a besoin des T° à 9h et à 15h pour mieux prévoir le maximum d'ozone. L'effet de la T° n'est pas exactement le même entre 9h et 15h (mais 12h n'est pas utile comme info). L'effet de la nébulosité sur le maximum d'O3 est le même quelle que soit la direction du vent puisque l'interaction nébulosité-vent n'est pas significative. De même, l'effet de la vitesse du vent sur le maximum d'O3 est le même quelle que soit la direction du vent puisque toutes les interactions entre la direction et les vitesses de vent sont non significatives. On peut évidemment continuer à interpréter la présence ou l'absence de tous les effets.

Diapositive 33 :

Après avoir commenté les effets globaux de chaque variable ou interaction, on peut interpréter plus dans le détail les coefficients non nuls. Le coefficient pour la T° à 9h vaut 1.82, ce qui n'est pas significativement différent de 0 (on l'avait vu dans la diapositive précédente, l'effet de la T° à 9h est présent à travers l'interaction avec la direction du vent). On ne va donc pas trop commenter ce coefficient. De même pour la T° à 15h.

Le coefficient de la nébulosité à 12h est égal à -3.01. Cela signifie que plus la nébulosité à 12h est grande, plus le maximum d'ozone a tendance à être petit. Pour éviter de mal interpréter, on vérifie que le signe du coefficient est le même que le signe du coefficient de corrélation entre la nébulosité et le maximum d'ozone. Plus précisément, le modèle dit que si la nébulosité à 12h augmente d'1 unité, toutes les autres variables de température, vitesse du vent, maximum d'ozone de la veille, restant égales par ailleurs, le maximum d'ozone diminue de -3. En pratique, il est souvent difficile de dire « toutes les autres variables restant égales par ailleurs » car il existe des liaisons entre les variables explicatives, et on ne peut donc pas voir ce que serait une augmentation d'une unité pour une variable, sans que les autres variables bougent.

Le coefficient du maximum d'ozone de la veille vaut +0.34 et donc plus le maximum d'O3 de la veille est grand, plus le maximum d'O3 est grand. Pour les 4 lignes suivantes, nous avons les coefficients de direction du vent.

On peut lire que les jours de vents d'est et surtout du nord ont des max d'O3 supérieur aux jours de vents d'ouest et du sud. Cependant, on l'a vu, l'effet seul de la direction du vent est non significatif, donc on ne peut pas affirmer que ces coefficients soient différents de 0. On ne va donc pas trop les commenter. Nous avons ensuite 4 coefficients pour l'interaction direction du vent et T° à 9h. Les jours de vent du nord, l'effet de la T° à 9h est plus important car le coef vaut 6.14. Ainsi, quand le vent vient du nord, l'effet de la T° à 9h est particulièrement fort (donc s'il fait chaud à 9h quand le vent vient du nord, le max d'O3 risque d'être important). Pour 1° de plus avec une T° à 9h, le maximum d'ozone augmente de 1.82 (ça c'est l'augmentation moyenne quelle que soit la direction du vent) + 6.14 = 7.96. Donc 1° de plus à 9h si le vent vient du nord, fait augmenter le maximum d'ozone de 7.96 (là encore, si toutes les autres variables restent constantes). Quand le vent vient de l'est, l'effet de la température à 9h est de 1.82-4.74= -2.92. Donc

si la T° à 9h augmente de 1° et que le vent vient de l'est, alors le maximum d'ozone diminue de 2.92, toujours si toutes les autres variables restent constantes. Cela peut paraître surprenant que le maximum d'ozone diminue si la T° à 9h est plus élevée. Mais en fait, il faut bien avoir en tête que ce serait le cas si toutes les autres variables restaient constantes. Or si la T° à 9h augmente de 1°, il y a de grande chance que la T° à 15h augmente aussi, et si la T° à 15h augmente, le coefficient entre vent d'est et T15 vaut 3.68, ce qui a tendance à augmenter le maximum d'ozone. C'est toute la difficulté des interprétations des coefficients quand les variables explicatives sont corrélées. Et c'est donc tout l'intérêt de pouvoir construire des plans d'expériences quand on peut choisir les expériences que l'on effectue.

Diapositive 34 :

Revenons sur la démarche à adopter en modélisation. Tout d'abord, on va lister toutes les variables qui entrent en jeu pour expliquer ou prédire la variable réponse. On n'hésitera pas à mettre plus de variables que pas assez. Ensuite, il est important de visualiser les données et notamment les liaisons avec la variable réponse pour s'assurer qu'il n'y a pas de valeurs aberrantes et pour avoir quelques idées sur les liaisons entre variables. Puis on écrit et construit le modèle en choisissant à la fois les effets et les interactions qui peuvent expliquer la réponse. On ne mettra pas systématiquement, et sans réfléchir, toutes les interactions possibles dans le modèle. On peut a priori écarter certaines interactions quand on connaît un peu le phénomène que l'on étudie. Et surtout, on évite de mettre des interactions d'ordre supérieur ou égal à 3. Les interactions pour lesquelles on ne peut pas décider a priori seront mises dans le modèle, quitte à construire un modèle avec beaucoup d'effets. Et on laissera les tests statistiques faire la sélection. Pour cela, à partir du modèle complet, on sélectionne le sous-modèle qui minimise l'AIC (ou le BIC ou à la main) en supprimant les interactions et effets non utiles, puis on construit ce sous-modèle.

Ensuite, et c'est le plus important, on interprétera les résultats, les effets significatifs comme ceux non significatifs, on interprétera les coefficients du modèle en faisant bien attention aux confusions possibles notamment pour les variables quantitatives, et enfin on peut utiliser le sous-modèle pour prédire de nouvelles valeurs si on a un objectif de prévision.