***Analyser le code la figure 2 ?***

***Comparer les 2 modèles ?***

**Etude sur les variantes d’une graminée**

Nous utilisons un jeu de données, portant sur l’assimilation du CO2 par une espèce de graminée. Cela nous permettra de voir comment faire une étude d’ANOVA, réaliser un modèle de régression statistique, prédire l’origine des plantes et de générer un nouveau jeu de données.

**1°) les données**

Le jeu de données est natif. Nous utilisons donc la commande data(CO2) afin de le charger en mémoire. Les données contenues sont l’assimilation du CO2, la concentration en CO2, la température, et la localisation. Cette dernière prend, soit la valeur Mississipi, soit la valeur Québec. C’est une variable qualitative. En revanche, l’assimilation, la concentration et la température sont des données quantitatives. Nous allons chercher les relations entre ces données.

**2°) analyse graphique**

Nous utilisons le code TIS4\_TD1\_2014-2015\_CodeGraph1.R, afin de générer le graphique correspondant à la figure 1.

Ce graphe montre pour le Québec et le Mississipi, la variation de l’assimilation du CO2 et la température ambiante et nocturne, en fonction de la concentration en CO2.

L’assimilation du CO2 et les températures augmentent avec la concentration en CO2, pour le Québec et pour le Mississipi. Egalement la température de nuit est inférieure à la température ambiante. Seulement, la température de nuit est beaucoup plus inférieure à la température ambiante pour le Mississipi que pour le Québec. Il est donc faisable de penser qu’il fait beaucoup plus froid la nuit que le jour au Mississipi, et qu’au Québec il fait tout le temps froid. Pour ce dernier les valeurs de l’absorption sont beaucoup moins disparates et atteignent 40 mmol/ms. En revanche les valeurs pour le Mississipi sont beaucoup hétérogènes et atteignent un maximum de 35 mmol/ms.

Par comparaison des cas du Québec et du Mississipi, on comprend que selon la localisation et la température, l’assimilation en C02 est différente. Il y doit y avoir un lien entre les données.

**3°) régression linéaire multiple et ANOVA**

Nous réalisons un modèle complet, expliquant l’assimilation du CO2 par les plantes.

L’**AIC** ou l’Akaike Information Criterion, retient le critère le plus faible. Il calcul le nombre d’erreur. Plus il est bas, plus le test est performant. Il peut être utilisé à la place de la cross-validation, qui prend beaucoup de temps en calcul.

D’autre part, le coefficient de détermination R^2, donne la qualité de l’ajustement. Plus R^2 tend vers 1, meilleur est la qualité d’ajustement. Le coefficient de détermination multiple, s’appuie sur l’équation de l’analyse des variances. Il permet d’évaluer globalement la pertinence du modèle de prédiction. Ce R^2 augmente lorsqu’il y a un grand nombre de variables. Dans ce cas mieux vaut utiliser le coefficient de détermination ajusté. Celui-ci est corrigé par les degrés de liberté associés aux variances. Nous pouvons utiliser le R^2, afin de déterminer si une variance expliquée entre 2 variables est significative ou pas. *Le R^2 multiple est de 0.7049 et le R^2 ajusté est de 0.6899. De plus, la p-valeur est de 2.10^-16. Cette p-valeur est basse. Elle indique, qu’il faut rejeter l’hypothèse H0. Il semble donc y avoir une variance explicative significative entre la variable d’état et l’assimilation. Aussi le modèle est explicatif.*

Nous procédons à un test d’Anova de type 2. Pour ce faire nous utilisons, la commande Anova(CO2.modele.complet). L’**ANOVA** est un cas particulier de la régression linéaire. Elle mesure entre autre, la force d’association entre une variable dépendante et une variable explicative. Le but est de trouver une source de variation parmi plusieurs facteurs de traitements. En effet, dans le cas des graminées nous essayons de montrer l’origine de la différence d’assimilation du C02 entre le Québec et le Mississipi. L’ANOVA de type 1 est séquentielle. En conséquence « on “pompe” la variance au fur et à mesure qu'on “résoud” chacun des facteurs. » Mieux vaut alors utiliser l’ANOVA de type 2 qui n’est pas séquentielle. *Les sommes des carrés observables à l’issue de ce test sont relativement grandes. Cela conforte l’idée précédemment émise, selon laquelle il a un lien entre la variable d’état et l’assimilation.*

Nous voulons établir un modèle de prédiction.

Nous générons un nouveau jeu de données. Nous utilisons le code TIS4\_TD1\_2014-2015\_CodeGraph2.R, afin de générer le graphe.

***Analyser le code la figure 2*?**

***Réalisation d’un modèle de régression linéaire de la variable assimilation***

Lequel illustre l’assimilation prévu, en fonction de l’assimilation observée. Cette méthode tente de prédire quelle quantité de CO2 devrait être assimilée pour chaque plante. Elle compare les données prédites et les données observées. Ainsi, elle donne une idée de la justesse du modèle. Le graphe montre la droite x=y ou droite de régression. Laquelle représente la prédiction parfaite. Plus les points sont proches de cette droite, meilleure est la prédiction. Le nuage de points est localisé autours la droite x=y. La prédiction semble donc, relativement correcte. Certains points sont presque localisés sur la droite. La prédiction est alors très performante dans ces cas-là. Certains points sont situés sur une droite parallèle à la droite de régression. Cela signifie que l’évolution de ces données est proportionnelle à la droite de régression. Il est alors possible de chercher une source d’erreurs.

Enfin nous réduisons le modèle en nous servant de la p-valeur associée à l’ANOVA. Nous nommons ce dernier modèle CO2.model. Nous calculons la valeur de l’AIC. Pour ce faire nous utilisons la commande AIC(m).

Au vu des résultats de l'ANOVA et des tests sur les coefficients pour les modèles complet et réduit, nous pouvons raisonnablement réduire le modèle complet.

En effet, quelle que soit l’ANOVA utilisée, il y a 2 variables qui sont significativement corrélées. Elles partagent par conséquent la même variance. Puisqu’elles se réfèrent à la même variabilité, elles peuvent être considérées interchangeables. L’une d’entre elles pouvant être ignorée au bénéfice de l’autre.

Au vu des AIC et R^2 respectifs des 2 modèles, ce nouveau modèle sans la variable

Plant / ID\_Plante semble être intéressant. En effet, les coefficients de détermination multiples et ajustés sont semblables pour les 2 modèles. Aussi le modèle réduit est intéressant car avec moins de variables il permet d’aboutir à un R^2 et un AIC comparables à ceux du modèle complet.

**4°) Régression logistique statistique**

Nous réalisons un modèle complet, que nous appelons Plante.model. Lequel prédit l’origine de la plante en fonction de l’assimilation et du traitement. Nous réduisons ce modèle au moyen de la commande step (model). Nous faisons également un modèle réduit de l’origine de la plante en fonction de l’interaction assimilation/traitement.

***Comparer les 2 modèles***

Enfin nous traçons la figure 3 grâce au code fournis : TIS4\_TD1\_2014-2015\_CodeGraph3.R

Ce dernier illustre la probabilité d’appartenir à un style de plante en fonction de l’assimilation au CO2. Nous constatons que plus l’assimilation de CO2 augmente, plus les courbent décroisent. Cela signifie que la probabilité d’appartenir au type Mississipi diminue quand l’assimilation de CO2 augmente. Ainsi, plus l’assimilation augmente, plus la probabilité que la plante soit du Québec augmente.

*Aussi ce modèle semble converger dans la même direction que les tests précédents. Il semble donc être efficace.*