TP n°2 – Prévision de volume de bois –

Pour réaliser ce TP **en binôme**, on utilisera le fichier de commandes R '**ScriptTP2.R'** (disponible sur **Campus**) ainsi que le fichier de données '**arbres.txt'**.

1. Le problème et sa modélisation

Un exploitant forestier contacte votre bureau d'études spécialisé en « data science ». Il souhaite estimer la quantité de bois d'une forêt arrivée à maturité de manière à fixer un prix de revient du bois. Pour cela, il est en particulier nécessaire de connaître la hauteur des arbres afin de calculer le volume par une formule du type « tronc de cône » qui consiste à assimiler chaque arbre à un cône circulaire droit (cône de révolution). L'exploitant fournit des données de **hauteur (en m)** et de **circonférence (en cm)** mesurée à **1m30** du sol (fichier **arbres.txt** sur **Campus**). Ces données correspondent à un échantillon d'arbres de taille **100** (voir figure ci-dessous). Le **nombre total** d'arbres de la forêt est estimé à **100,000**. En pratique, la hauteur d'un arbre en pleine forêt est difficile à calculer à moins de l'abattre. Par contre, l'exploitant forestier mesure facilement la circonférence et peut donc en connaître sa distribution. Il vous précise que cette distribution est très bien approchée par une loi normale $\mathbf{N(\mu_C, \sigma_C^2)}$ avec μ_C **= 47 cm** et σ_C **= 8.5 cm**.

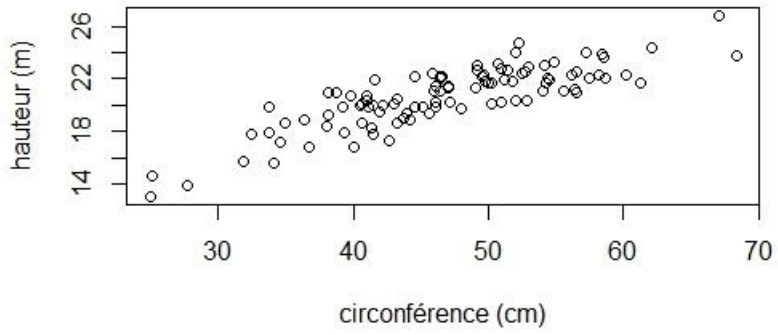

Figure. Nuage d'arbres décrits par leur circonférence et hauteur

Pour estimer la quantité totale de bois (en m^3), on propose d'utiliser un modèle de régression linéaire pour apprendre la relation entre la hauteur et la circonférence d'un arbre. Ce modèle sera ensuite utilisé pour construire un intervalle de prédiction unilatéral à 95% du volume total de bois.

2. Une première « fausse » prédiction à 95% du volume total de bois

Cela consiste à supposer que l'échantillon est représentatif de la forêt, i.e. un échantillon au sens statistique du couple de variables (*circonférence*, *hauteur*) qui décrit un arbre. Sous cette hypothèse, on peut calculer les volumes individuels correspondants v_1, …, v_n, ce qui constitue un échantillon statistique de la variable *volume* associée à un arbre. Soit maintenant la variable *volume total de bois* de la forêt $V_T = V_1 + \ldots + V_{Na}$ avec N_a =100,000 arbres et V_k volume de l'arbre n° k.

D'après le théorème de la limite centrée (sous réserve que les variables V_k soient indépendantes), on sait que V_T à peu près de loi normale $N(N_a \times \mu_V ; N_a \times \sigma_V^2)$, ce qui permet d'estimer le quantile à 5% de V_T par la formule $q_{5\%}(V_T) \approx Na \times \mu_V - 1.64 \times \sqrt{N_a} \times \sigma_V$. C'est une première estimation, qui est bien sûr mauvaise si l'échantillon d'arbres n'est pas du tout représentatif de la forêt. C'est à cause de cette hypothèse très restrictive que l'on va maintenant développer un modèle plus élaboré.

3. Un modèle de régression linéaire simple

En notant *y* la variable « hauteur » et *x* la variable « circonférence », on propose le modèle de régression linéaire simple suivant :

$$y = \beta_0 + \beta_1 x + \varepsilon$$

3.1 Estimation et prédictions. Estimer les paramètres du modèle. Commentaires. Obtenir un graphique montrant la réponse moyenne estimée et l'intervalle de confiance à 95% en fonction de la valeur de *x*. Superposer l'intervalle de prédiction à 95%.

3.2 Une première prévision de volume total de bois. Pour estimer le volume total de bois, on propose une méthode de Monte-Carlo.

- Dans un premier temps, écrire une fonction R qui permette de simuler une forêt de 100,000 arbres décrits par le couple (hauteur, circonférence). Pour cela, on commence par simuler les 100,000 valeurs de circonférence en utilisant la loi de la variable circonférence. Ensuite, on utilise le modèle linéaire pour simuler la hauteur correspondante des arbres.

- Dans un second temps, écrire une fonction qui calcule le volume total de bois (en m^3) d'une forêt composé de 100,000 arbres en utilisant la formule qui donne le volume d'un cône de révolution connaissant sa circonférence à 1m30 du sol et sa hauteur totale.

On est donc en mesure de simuler une forêt de taille 100,000 et le *volume total de bois* correspondant. Pour construire l'intervalle de prévision à 95% demandé par l'exploitant, on propose tout simplement d'obtenir N réalisations indépendantes de la variable *volume total de bois* puis de calculer le quantile empirique d'ordre 5% de ces N réalisations. Le nombre N désigne le nombre de simulations Monte-Carlo utilisées. On pourra tracer le quantile empirique d'ordre 5% en fonction de N et choisir la taille convenable du nombre N à partir de la courbe obtenue.

3.3 Correction pour tenir compte de l'incertitude sur les paramètres du modèle

En quoi la technique Monte-Carlo utilisée présente un biais ? Comment faire pour corriger ce biais. Mettre en œuvre cette solution et comparer le résultat obtenu avec ce qui précède.

3.4 Analyse des résidus et validation du modèle de régression linéaire simple. Faire l'analyse des résidus. En particulier, obtenir un graphique des résidus « studentisés » contre la réponse prédite. Validez-vous le modèle ?

4. Un modèle de régression linéaire multiple. A la lumière de la figure montrant le nuage de points associés aux arbres, proposer un modèle de régression linéaire multiple en introduisant une variable supplémentaire bien choisie qui soit une fonction simple de la circonférence. Reprendre toutes les étapes de la partie précédente et comparer les intervalles de prévision obtenus. Vérifier enfin que vous validez bien le modèle…