– Texte –
La hauteur des eucalyptus

Mots-clefs : Régression linéaire simple, régression linéaire multiple, prévision.

1 Le problème et sa modélisation

Lorsqu'on cherche à estimer la quantité de bois produite par une forêt, il est nécessaire de connaître la hauteur des arbres afin de calculer le volume par une formule de type "tronc de cône". Cependant, mesurer la hauteur d'un arbre d'une vingtaine de mètres n'est pas chose facile : on utilise en général un dendromètre, lequel mesure un angle entre le sol et le sommet de l'arbre et nécessite donc une vision claire de la cime ainsi qu'un recul assez grand pour avoir une mesure précise de l'angle.

Lorsque ces conditions ne sont pas réunies, on peut chercher à estimer cette hauteur via un modèle de régression linéaire à partir de la simple mesure de la circonférence à 1 mètre 30 du sol. Cette modélisation nécessite un échantillon d'apprentissage, c'est-à-dire un ensemble d'arbres pour lesquels ont été réellement mesurées la circonférence et la hauteur. La figure 1 représente un nuage de points pour des mesures effectuées sur environ 1400 eucalyptus.

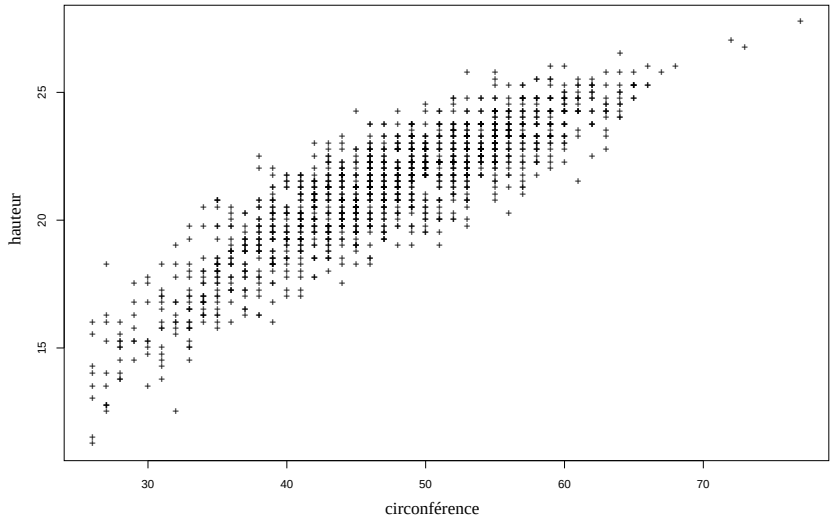

Figure 1 – Nuage de points pour les eucalyptus.

2 Régression linéaire simple

Si on note X la circonférence d'un arbre à 1 mètre 30 du sol et Y sa hauteur, le modèle de régression linéaire simple revient à supposer une relation de la forme :

$$Y = \beta_1 + \beta_2 X + \varepsilon.$$

En d'autres termes, on considère que la hauteur dépend linéairement de la circonférence, mais que cette liaison est perturbée par une erreur. Dans un tel modèle, X est appelée variable explicative tandis que Y est la variable à expliquer.

On dispose de n observations de la circonférence X, notées $(x_i)_{1\leqslant i\leqslant n}$, pour lesquelles on connaît les hauteurs respectives $(y_i)_{1\leqslant i\leqslant n}$. Quitte à noter les variables aléatoires en minuscules, ceci donne :

$$\forall i \in \{1,\dots,n\} \qquad y_i = \beta_1 + \beta_2 x_i + \varepsilon_i,$$

où les x_i sont connus (donc non aléatoires), les paramètres β_1 et β_2 sont inconnus et à estimer, les ε_i sont les réalisations inconnues d'une variable aléatoire ε et les y_i sont les observations de variables aléatoires.

Exemple : Sur le site `http://w3.bretagne.ens-cachan.fr/math/OrauxBlancs/`, importer les fichiers `circ.txt` et `h.txt` dans le répertoire où vous ouvrez Scilab, puis taper : `>x=fscanfMat('circ.txt'); y=fscanfMat('h.txt');`

Définition 1 : Estimateurs des moindres carrés
On appelle estimateurs des moindres carrés de β_1 et β_2 les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ obtenus par la minimisation suivante :

$$(\hat{\beta}_1, \hat{\beta}_2) = \arg\min_{(\beta_1,\beta_2)} S(\beta_1,\beta_2) = \arg\min_{(\beta_1,\beta_2)} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2.$$

Les moyennes empiriques des circonférences et des hauteurs sont respectivement notées $\bar{x}$ et $\bar{y}$, c'est-à-dire $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. On fait l'hypothèse (naturelle) suivante :

Hypothèse $(\mathcal{H}_1)$ **:** Il existe i et j tels que $x_i \neq x_j$.

Ceci supposé, les estimateurs s'expriment facilement en fonction des observations :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \& \qquad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

Pour pouvoir établir des propriétés de biais et de variance des estimateurs des moindres carrés, il faut bien sûr une hypothèse sur les erreurs ε_i.

Hypothèse $(\mathcal{H}_2)$ **:** *Les erreurs ε_i sont centrées, de même variance σ^2 (homoscédasticité) et décorrélées entre elles.*

Propriété 1 : Biais, Variances et Covariance
Sous les hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$:
- *Les estimateurs sont sans biais : $\mathbb{E}[\hat{\beta}_1] = \beta_1$ et $\mathbb{E}[\hat{\beta}_2] = \beta_2$.*
- *Les variances et covariance sont :*

$$\mathrm{V}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}, \mathrm{V}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}.$$

Preuve. L'expression suivante peut s'avérer commode pour démontrer certains points

$$\hat{\beta}_2 = \beta_2 + \frac{\sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2}.$$

■

Pour avoir une idée des variances et covariance de $\hat{\beta}_1$ et $\hat{\beta}_2$, les formules de la Propriété 1 ne sont pas pratiques car elles font intervenir la variance σ^2 de l'erreur, qui est généralement inconnue. Il faut donc l'estimer elle aussi. La preuve du résultat suivant sera vue plus loin.

Propriété 2 : Estimateur de la variance
Un estimateur sans biais de σ^2 est :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2}{n-2}.$$

3 Régression linéaire multiple

Le nuage de points de la figure 1 peut laisser penser, notamment pour les petites valeurs de la circonférence, qu'une modélisation incluant la racine carrée de celle-ci pourrait être judicieuse. C'est ce que nous allons faire maintenant, en généralisant la méthode de la section précédente. Nous supposons donc cette fois que pour tout $i \in \{1, \ldots, n\}$:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 \sqrt{x_i} + \varepsilon_i,$$

et le but est d'estimer β_1, β_2 et β_3. Passons en notations matricielles :

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_1 & \sqrt{x_1} \\ \vdots & \vdots & \vdots \\ 1 & x_n & \sqrt{x_n} \end{bmatrix}, \; \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

Ainsi l'estimateur des moindres carrés $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]^T$ s'écrit tout simplement :

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2,$$

en notant $\|.\|$ la norme euclidienne de $\mathbb{R}^n$.

Proposition 2 : Estimateur des moindres carrés
Sous l'hypothèse $(\mathcal{H}_1)$, l'estimateur des moindres carrés est : $\hat{\beta} = (X^TX)^{-1}X^TY$.

Preuve. Si $\mathcal{F}$ est le sous-espace de $\mathbb{R}^n$ engendré par les vecteurs colonnes de X, tout vecteur de $\mathcal{F}$ est de la forme $X\beta$. Ainsi la quantité $\|Y - X\beta\|^2$ est minimale lorsque $X\beta$ correspond au projeté orthogonal de Y sur $\mathcal{F}$, projeté que l'on note donc $X\hat{\beta}$. Ce projeté est caractérisé par le fait que pour tout vecteur α, $\langle X\alpha, Y - X\hat{\beta}\rangle = 0$, d'où l'on déduit bien $\hat{\beta} = (X^TX)^{-1}X^TY$. ■

Remarque : Ceci signifie simplement que $P_X = X(X^TX)^{-1}X^T$ est la matrice de projection orthogonale sur $\mathcal{F}$. En notant $P_{X^\perp} = (I_n - P_X)$ la matrice de projection orthogonale sur $\mathcal{F}^\perp$, le minimum de S est donc $S(\hat{\beta}) = \|P_{X^\perp}Y\|^2$.

Les résultats vus en section précédente se généralisent alors sans problème.

Propriété 3 : Biais, Dispersion et Estimation de σ^2
Sous les hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$:
- *L'estimateur $\hat{\beta}$ est sans biais.*
- *La matrice de covariance de $\hat{\beta}$ est : $\Gamma = \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = \sigma^2(X^TX)^{-1}$.*
- *L'estimateur $\hat{\sigma}^2 = \|Y - X\hat{\beta}\|^2/(n-3)$ est un estimateur sans biais de σ^2.*

Preuve. Pour l'estimateur de σ^2, on fait intervenir la trace :

$$\mathbb{E}[\|Y - X\hat{\beta}\|^2] = \mathbb{E}[\|P_{X^\perp}Y\|^2] = \mathbb{E}[\|P_{X^\perp}\boldsymbol{\varepsilon}\|^2] = \mathbb{E}[\mathrm{Tr}(\|P_{X^\perp}\boldsymbol{\varepsilon}\|^2)],$$

c'est-à-dire :

$$\mathbb{E}[\|Y - X\hat{\beta}\|^2] = \mathrm{Tr}(\mathbb{E}[P_{X^\perp}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T P_{X^\perp}]) = \mathrm{Tr}(P_{X^\perp}\sigma^2 P_{X^\perp}) = \mathrm{Tr}(P_{X^\perp})\sigma^2 = (n-3)\sigma^2,$$

puisque $P_{X^\perp}$ est la matrice d'une projection sur un sous-espace de dimension $(n-3)$. ■

4 Modèle gaussien

On voudrait maintenant savoir si l'ajout de la racine carrée dans le modèle est vraiment pertinent. Ceci peut se faire facilement dans le cas où les erreurs sont supposées gaussiennes, hypothèse que nous ferons dans toute la suite.

Hypothèse $(\mathcal{H}_3)$: Les erreurs ε_i sont indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$.

4.1 Estimateur du maximum de vraisemblance

En généralisant un peu les notations de la section 3, on se place dans le cadre d'un problème de régression linéaire à p paramètres :

$$Y = X\beta + \boldsymbol{\varepsilon},$$

où Y est un vecteur colonne $n \times 1$ de terme générique y_i, X une matrice $n \times p$ de terme générique x_{ij} et dont la première colonne est constituée de 1, β est un vecteur colonne $p \times 1$ de terme générique β_j et $\boldsymbol{\varepsilon}$ est un vecteur colonne $n \times 1$ de terme générique ε_i. La vraisemblance de l'échantillon en fonction des paramètres β et σ^2 s'écrit :

$$\mathcal{L}_n(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right).$$

Si on note $(\hat{\beta}_{mv}, \hat{\sigma}^2_{mv})$ les estimateurs au maximum de vraisemblance et que l'on conserve les notations $(\hat{\beta}, \hat{\sigma}^2)$ pour les estimateurs des moindres carrés, on a des liens très simples entre ces quantités.

Proposition 3 : Moindres carrés et maximum de vraisemblance
Sous les hypothèses $(\mathcal{H}_1)$ *et* $(\mathcal{H}_3)$*, on a* $\hat{\beta}_{mv} = \hat{\beta}$ *et* $\hat{\sigma}^2_{mv} = (n-p)\hat{\sigma}^2/n$*. De plus,* $\hat{\beta}_{mv}$ *et* $\hat{\sigma}^2_{mv}$ *sont indépendants.*

Preuve. Pour l'indépendance, il suffit de remarquer que $\hat{\beta} = (X^TX)^{-1}X^TP_XY$ tandis que $\hat{\sigma}^2_{mv} = \|P_{X^\perp}Y\|^2/n$. Or les vecteurs gaussiens P_XY et $P_{X^\perp}Y$ sont des projections sur des sous-espaces orthogonaux, donc ils sont décorrélés, donc indépendants.

■

4.2 Test de Student

La loi de Student à d degrés de liberté, notée $\mathcal{T}_d$, fait intervenir le rapport indépendant entre une variable normale centrée réduite et la racine carrée d'un chi-deux à d degrés de liberté. Pour aller vite :

$$\mathcal{T}_d = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_d/d}}.$$

Sa fonction quantile est notée t_d, c'est-à-dire que si $T \sim \mathcal{T}_d$, on a : $\alpha = \mathbb{P}(T \leqslant t_d(\alpha))$.

Supposons comme en section 4.1 un modèle à p variables satisfaisant les hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_3)$: $Y = X\beta + \boldsymbol{\varepsilon}$. On veut tester la nullité du dernier coefficient β_p de β :

$$\mathrm{H}_0 \ : \ \beta_p = 0 \qquad\qquad \mathrm{H}_1 \ : \ \beta_p \neq 0,$$

test bilatéral de significativité de β_p. On effectue pour cela un test de Student avec la statistique de test $T = \hat{\beta}_p / \hat{\sigma}_{\hat{\beta}_p}$, où :

$$\hat{\sigma}_{\hat{\beta}_p} = \hat{\sigma}\sqrt{((X^TX)^{-1})_{p,p}} = \frac{\sqrt{((X^TX)^{-1})_{p,p}}}{\sqrt{n-p}}\|Y - X\hat{\beta}\|.$$

La variable aléatoire T suit sous H_0 une loi de Student à $(n-p)$ degrés de liberté. On rejette donc H_0 au niveau de confiance $1-\alpha$ si l'observation sur notre échantillon de la statistique T, notée $T(\omega)$, est telle que $|T(\omega)| > t_{n-p}(1-\alpha/2)$. Sur l'exemple précédent, ceci nous amène à conserver la racine carrée dans le modèle.

4.3 Prévision

Pour une nouvelle valeur x_{n+1} de la circonférence, nous voulons prédire la hauteur y_{n+1}. En notant $x'_{n+1} = [1, x_{n+1}, \sqrt{x_{n+1}}]$, le modèle nous dit que $y_{n+1} = x'_{n+1}\beta + \varepsilon_{n+1}$. Grâce à l'estimateur $\hat{\beta}$ obtenu en section 3, nous prévoyons y_{n+1} par $\hat{y}^p_{n+1} = x'_{n+1}\hat{\beta}$. L'erreur de prévision est alors définie par $\varepsilon^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$, dont on voit aisément qu'elle suit une loi normale centrée de variance

$$\mathrm{V}(\varepsilon^p_{n+1}) = \sigma^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1}).$$

Nous pouvons en déduire un intervalle de confiance de niveau $(1-\alpha)$ pour y_{n+1} :

$$IC(y_{n+1}) = \left[x'_{n+1}\hat{\beta} \pm t_{n-3}(1-\alpha/2)\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}\right].$$

Cet intervalle de confiance est d'autant plus grand que x_{n+1} est loin de la moyenne empirique $\bar{x}$.

5 Suggestions

- Démontrer la Propriété 1.
- Représenter sur un même graphe le nuage de points, la droite de régression et la courbe obtenue en tenant compte de la racine carrée de la circonférence.
- Démontrer la Propriété 3.
- Démontrer la Proposition 3.
- Expliquer pourquoi le test proposé est bien un test de Student.
- Effectuer le test sur l'exemple des eucalyptus.
- Représenter les courbes correspondant aux intervalles de confiance de prévision de la section 4.3.

