M2 recherche fiche 5: régression linéaire et orthogonalisation

Emeline Schmisser, `emeline.schmisser@math.univ-lille1.fr`, bureau 314 (bâtiment M3).

1 Modèle

On observe des variables $y_1, \ldots, y_n$, $x_{0,1}, \ldots, x_{0,n}$, ..., $x_{m,1}, \ldots, x_{m,n}$.
On pense que les variables y_i suivent le modèle

$$Y = a_0 X_0 + a_1 X_1 + \ldots + a_m X_m + \varepsilon$$

Ce sont des notations générales, on peut avoir $X_0 = 1$,$X_2 = X_1^2$ ou $X_3 = \sin(X_1)$. On note

$$Y = (y_1, \ldots, y_n)^* \quad X = (X_0, X_1, \ldots, X_m) \quad A = (a_0, a_1, \ldots, a_m)^* \quad X_i = (x_{i,1}, \ldots, x_{i,n})^*$$

2 Principe de la régression linéaire

On veut minimiser l'expression

$$\frac{1}{n} \sum_{k=1}^{n} (y_k - (a_0 x_{0,k} + a_1 x_{1,k} + \ldots + a_m x_{m,k}))^2$$

Cela revient à minimiser la distance

$$\|Y - XA\|^2 = \|Y - (a_0 X_0 + a_1 X_1 + \ldots + a_m X_m)\|^2$$

Notons $\hat{A}$ le vecteur qui minimise l'équation précédente. $X\hat{A}$ est le projeté orthogonal de Y sur le sous-espace vectoriel $Vect(X_0, X_1, \ldots, X_m)$. Le vecteur $\hat{\varepsilon} = Y - X\hat{A}$ est orthogonal au sous-espace vectoriel $Vect(X_0, X_1, \ldots, X_m)$ et donc, pour tout $i \in [0, \ldots, m]$,

$$Cov(X_i, \hat{\varepsilon}) = \langle X_i, \hat{\varepsilon} \rangle / n = 0.$$

Exercice 1 (premières questions)

1. Expliquer ce qui se passe si la matrice X n'est pas de rang m. Dans la suite, on suppose toujours que le rang de X est m.
2. On suppose que $X_0 = 1$ (X_0 constant). Que se passe-t-il si on recentre toutes les colonnes sauf la première ? Est-ce qu'on change l'espace de projection ? Comment sont modifiés les coefficients ?
3. Que vaut XX^* dans ce cas ?

Exercice 2 (La régression comme projection linéaire)

1. Montrer que le vecteur $X(X^*X)^{-1}X^*$ est bien défini.
2. Montrer que $X(X^*X)^{-1}X^*B$ est le projeté orthogonal du vecteur B sur le sous-espace vectoriel $Vect(X_0, X_1, \ldots, X_m)$. En déduire l'expression de $\hat{A}$.

3. Que vaut $\hat{A}$ si on suppose $Y = aX + b$? Comparer avec les résultats obtenus précédemment.
4. Construire une fonction `regression2` qui, étant donné Y et la matrice X, renvoie le vecteur A. Appliquer cette fonction aux vecteurs $(y_1, \ldots, y_n)$, $(1, ..., 1)$, $(x_1, \ldots, x_n)$ construits lors du dernier TP. Attention au sens de construction de la matrice et à la commande pour multiplier les matrices ! Vérifier que l'on trouve bien les mêmes résultats que précédemment. Quelle est l'avantage de cette fonction par rapport à celle que l'on a construit précédemment ?

3 Loi de l'estimateur et test de nullité

Nous supposons maintenant que $Y = XA + \varepsilon$, avec ε_k des variables iid de loi $\mathcal{N}(0, \sigma^2)$.

Exercice 3 (loi de $\hat{A}$)

1. Quelle est la loi de $\hat{A}$?
2. On suppose σ^2 connu. Quelle est la loi du coefficient $\hat{a}_i$?
3. Que vaut $\hat{a}_i$ dans le cas où $Y = aX + b$? Est-ce que cet estimateur est convergent ?
4. Si σ^2 est connu, construire un test pour déterminer si le coefficient a_i est nul (sur papier). Attention : le fait que le coefficient soit nul ne veut pas forcément dire que X_i et Y ne sont pas corrélés, peut-être que Y et X_i dépendent tout les deux d'une autre variable.
5. Expliquer pourquoi quand $n \to \infty$, la probabilité de dire que a_i est nul alors que $|a_i| > \gamma$, avec γ fixé, tend vers 0. On démontrera le résultat dans le cas simple $Y = aX + b$.
   Dans le cas normal, on ne connaît pas σ^2. On l'estime :
   $$s^2 = \frac{1}{n - m - 1} \sum_{k=1}^{n} \hat{\varepsilon}_k^2$$
   $m + 1$ est le nombre de paramètres du modèle.
   La statistique s^2 suit une loi du χ^2 à $n - m - 1$ degrés de liberté, et :
   $$\frac{\hat{a}_i - a_i}{s\sqrt{(X^*X)^{-1}_{ii}}} \sim T(n - m - 1)$$
6. Contruire une fonction `estnul` qui étant donné Y, X et $\hat{A}$, renvoie un vecteur booléen (`TRUE` ou `FALSE`) suivant que le coefficient a_i est testé nul ou non (avec une erreur de type 1 à 10%).

4 Application

On reprend maintenant les variables x_i, y_i simulées lors de la dernière séance. On suppose que le modèle est $Y = a + bx + c\sin(x) + dx^2 + \varepsilon$.

Exercice 4 ()

1. Utiliser la fonction `regression2` pour faire la régression.
2. Utiliser R pour faire la régression. Comparer les résultats.
3. Tracer la courbe obtenue. Commenter.
4. Quelle est la moyenne des carrés des résidus ? Est-ce que cette moyenne a diminué par rapport au cas où on croyait que $Y = aX + b$?
5. Quelle est la covariance empirique entre :

- les résidus et x ?
- les résidus et $\sin(x)$?
- les résidus et x^2 ?
- les résidus et x^3 ?
- les résidus et $(x-1)^2$?

Quelle est la somme des résidus ?
Que vaut $\sum(x_k\hat{\varepsilon}_k)$?
Est-ce que ces résultats sont compatibles avec la théorie ?

6. Faire le test de nullité des coefficients. Que peut-on conclure ?
   On change maintenant le modèle : on considère $Y = a + bx + c\sin(x) + \varepsilon$.
7. Utiliser R pour faire la régression. Tracer la courbe obtenue.
8. Comparer les résultats de R avec ce que l'on obtient avec la fonction `regression2`.
9. Quelle est la variance empirique des résidus ? La moyenne des résidus ? Comment a varié la variance empirique des résidus par rapport à la situation précédente ?

5 Estimation adaptative : sélection du meilleur modèle

Si la variance empirique des résidus diminue, on peut penser que notre modèle est meilleur. C'est vrai si le nombre de coefficients est fixé, moins si il peut varier : on peut préférer un modèle avec moins de variables explicatives, donc plus simple à expliquer.

La variance empirique des résidus ne peut pas être un bon critère, car il n'est pas borné. À la place, on considère le critère de détermination simple

$$R^2 = 1 - \frac{Var(\hat{\varepsilon})}{Var(Y)}$$

Encore une fois, c'est un bon critère seulement si le nombre de paramètres est fixé. On introduit aussi le critère de détermination ajustée

$$\bar{R}^2 = 1 - \frac{Var(\hat{\varepsilon})/(n-p-1)}{Var(Y)/(n-1)}$$

Exercice 5 (Sélection du meilleur modèle)

1. Montrer que R^2 est compris entre 0 et 1. Que se passe-t-il si Y est complètement expliqué par X ? Si Y est indépendant de X ?
   Indication: : faire les calculs avec les vecteurs (c'est plus facile de comprendre ce qui se passe).
2. Construire une fonction qui calcule R^2 et $\bar{R}^2$ étant donné Y, X et $\hat{A}$.
3. Calculer R^2 et $\bar{R}^2$ dans les trois situations que l'on a rencontré. Conclure.