

COURS DE DATA MINING

10 : MODELISATION SUPERVISEE

REGRESSION LINEAIRE

7 séances de 3 heures

mai-juin 2007

EPF - 4^{ème} année - Option Ingénierie d’Affaires et de Projets

Bertrand LIAUDET

10 : Modélisation supervisée - 2 : Regression linéaire	2
Présentation	2
La régression linéaire simple	2
La régression linéaire multiple	3
Qualité de la régression : la question du résidu	3
Lecture des sorties de Clementine	6
Récapitulatif	6
Statistiques descriptives	6
Corrélations	6
Récapitulatif	6
Mesures de l’ANOVA	6
Statistiques des résidus	6
Conclusion : principales méthodes de modélisation qui n’ont pas été abordées.....	7
Classification	7
Prédiction	7

10 : MODELISATION SUPERVISEE

- 2 : REGRESSION LINEAIRE

Présentation

La régression linéaire simple

La régression linéaire simple permet de mettre en relation deux variables continues : la variable cible Y et la variable explicative X.

Quelles que soient les variables continues X et Y, on a :

$$y_i = a \cdot x_i + b + R_i$$

avec:

y_i : valeur de Y pour l'individu i

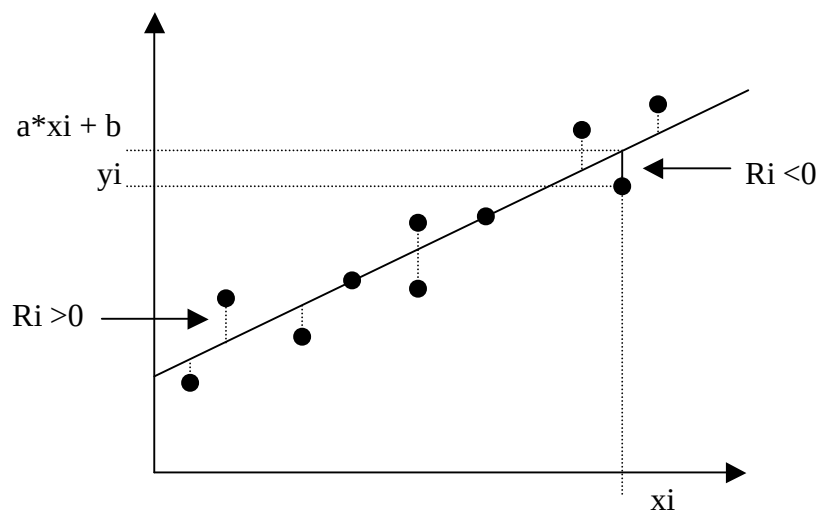
x_i : valeur de X pour l'individu i

a et b : coefficients de l'équation de régression linéaire

R_i : résidu pour l'individu i

La partie « $a \cdot x_i + b$ » est la composante déterministe du modèle.

La partie R_i est la composante stochastique appelée « erreur » ou « résidu ».



Remarque : des individus ayant la même valeur de X peuvent avoir des valeurs de Y différentes.

La droite $Y = aX + b$ est la droite de corrélation linéaire. On dit qu'elle « ajuste » le nuage de points.

La régression linéaire multiple

La régression linéaire multiple suit le même principe que celui de la régression linéaire simple. Elle permet de mettre en relation une variable continue cible Y et plusieurs variables continues explicatives X_k .

Quelles que soient les variables X_k et Y , on a :

$$y_i = a_1 \cdot x_{1i} + a_2 \cdot x_{2i} + a_3 \cdot x_{3i} + \dots + a_k \cdot x_{ki} + b + R_i$$

avec:

y_i : valeur de Y pour l'individu i

x_{ki} : valeur de X_k pour l'individu i

a_k et b : coefficients de l'équation de régression linéaire

R_i : résidu pour l'individu i

La partie « $a_1 \cdot x_{1i} + a_2 \cdot x_{2i} + a_3 \cdot x_{3i} + \dots + a_k \cdot x_{ki} + b$ » est la composante déterministe du modèle.

La partie E_i est la composante stochastique appelée « erreur » ou « résidu ».

Remarque : des individus ayant les mêmes valeurs de X_k peuvent avoir des valeurs de Y différentes.

Qualité de la régression : la question du résidu

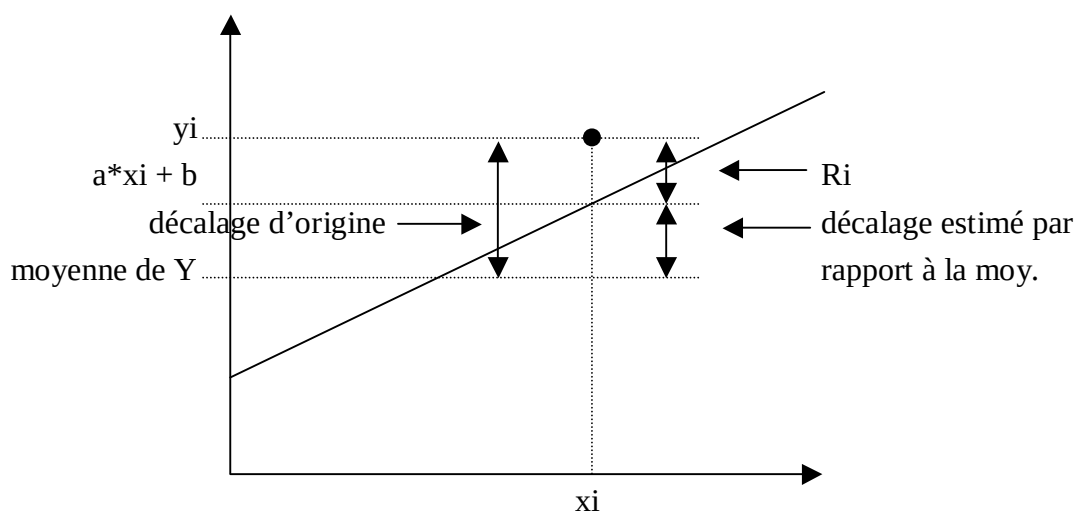
Technique

S_{Res} : somme des R_i^2 (somme des résidus au carré).

S_{Reg} : somme des carrés des décalages entre le y estimé et la moyenne de Y

$SDOM$: somme des carrés des décalages entre y_i d'origine et la moyenne de Y

Avec $SDOM = S_{Res} + S_{Reg}$



Méthode des moindres carrés

Elle consiste à chercher les coefficients « a » et « b » qui minimisent la somme des R_i^2 .

RMSE

RMSE : root mean square error, c'est la « moyenne » des carrés des résidus.

$$RMSE = S_{Res} / (n - p - 1)$$

Avec :

n : nombre d'individus

p : nombre de variables explicatives

(n-p-1) : nombre de degrés de liberté

La régression est d'autant meilleure que le **RMSE est petit**

Le F-ratio

$$F = n * S_{Reg} / S_{Res}$$

Cas des régressions multiples

$$F = (n-p-1) * S_{Reg} / (p * S_{Res})$$

Le R^2

$$R^2 = S_{Reg} / (S_{Reg} + S_{Res})$$

Dans le cas de la régression linéaire simple :

R = coefficient de corrélation de Pearson

La régression est d'autant meilleure que le **R^2 est proche de 1**

Le R^2 ajusté

Dans le cas des régressions multiples, on utilise le « R^2 ajusté » plutôt que le R^2 qui croit avec le nombre de variables, ce qui rend son interprétation biaisée.

$$R^2 \text{ ajusté} = 1 - (1 - R^2) * (n-1) / (n-p-1)$$

La régression est d'autant meilleure que le **R^2 ajusté est proche de 1**

Analyse de la variation des résidus

On analyse le nuage de points avec Y en axe horizontal et les résidus en axe vertical.

- Une répartition régulière des résidus doit donner une bande horizontale régulièrement répartie autour de 0.
- Autocorrélation des résidus. Si on tend à avoir une courbe et non plus une bande, cela veut dire que certaines valeurs seront sur-estimées tandis que d'autres seront sous-estimées.

L'autocorrélation se repère grâce au test de Durbin-Watson. Ce test calcule la somme des $(R_i - R_{i-1})^2 / \text{somme}(R_i)^2$. Cette valeur est comprise entre 0 et 4, < 2 pour des corrélations positives, > 2 pour des corrélations négatives.

Il faut un test de **Durbin-Watson compris entre 1,5 et 2,5** pour que l'autocorrélation soit acceptable.

- Autocorrélation des valeurs absolues des résidus. Si on tend à avoir un « cône » et non plus une bande horizontale, autrement dit, une augmentation ou une diminution régulière des résidus, il est recommandé de **remplacer les moindres carrés ordinaires par les moindres carrés pondérés**.

Ce « cône » se traduirait par une droite en faisant un nuage de points entre Y et la valeur absolue des résidus.



Lecture des sorties de Clementine

Récapitulatif

Il donne l'équation de régression linéaire

Statistiques descriptives

Moyenne et écart-type avec le nombre d'individus pris en compte (N).

Corrélations

Coefficient de Person = R

Récapitulatif

R, R^2 , R^2 ajusté, avec $R^2 = SReg / (SRes + SReg)$

Erreur standard de l'estimation = Racine ($SRes / ddl(Res)$) = Racine(RMSE)(cf. ANOVA).

Critère de Durbin-Watson.

Mesures de l'ANOVA

L'ANOVA, c'est l'analyse de la variance.

Somme des carrés des régressions : somme des carrés des décalages estimé par rapport à la moyenne de Y : SReg

Somme des carrés des résidus (on peut vérifier le résultat en faisant le calcul « à la main » avec Clementine) : SRes

Carré moyen des résidus : $SRes / ddl(res) = RMSE$

Carré moyen des régressions : $SReg / ddl(reg)$

F : $SReg * ddl(Res) / (SRes * ddl(Reg))$

ddl : degrés de liberté

Statistiques des résidus

Prévision et résidu.

Résidu min : pour voir de combien la prévision peut être décalée vers le bas.

Résidu max : pour voir de combien la prévision peut être décalée vers le haut.

Ecart-type du résidu : pour voir une sorte de résidu moyen (puisque la moyenne vaut 0). A comparer avec l'écart-type et l'amplitude de la prévision de la prévision.

Conclusion : principales méthodes de modélisation qui n'ont pas été abordées

Classification

Les réseaux de neurones

Prédiction

Analyse discriminante

Régression logistique

Réseaux de neurones