

Modèles Linéaires

TD 3 : Modèle Quadratique et comparaison de modèles Licence 3 MIASHS

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Résumé

Nous nous sommes précédemment intéressés au problème de la régression linéaire simple en se focalisant sur les estimateurs et leurs propriétés. Nous avons également effectué des tests statistiques et construit des intervalles de confiance sur ces derniers.

La présente fiche se présente comme une introduction au modèle multiple, *i.e.*, utilisant plusieurs variables indépendantes pour la prédictions des valeurs prises par une variable dépendante.

Plus précisément nous allons :

- introduire le modèle quadratique
- présenter des critères d'évaluations et de comparaisons de modèles
- comparer un modèle de régression linéaire simple avec un modèle de régression quadratique.

Estimation de la pureté d'un liquide

On cherche à établir un modèle permettant de prédire la pureté (notre variable Y) d'un liquide en fonction de la durée de filtration (notre variable X) de ce dernier. Pour cela on travaillera avec le jeu de données *purity*.

Nous allons également considérer deux modèles de régression

- le modèle de régression simple

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- le modèle de régression quadratique, qui est un cas particulier de modèle de régression multiple

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Construction et évaluation des modèles

Nous avons vu, en cours qu'un critère permettant d'évaluer la qualité d'un modèle s'appelle le **coefficient de détermination** R^2 .

1. Rappeller la définition du R^2 .
2. Construire votre modèle de régression linéaire simple. Dire si le modèle est globalement significatif et évaluer le R^2 de ce dernier.
*On utilisera la fonction **summary** de R pour déterminer ces éléments.*

```
summary([nom_du_modèle])
```

3. Faire de même avec le modèle de régression dit quadratique.
4. Les paramètres du modèle sont-ils significatifs ?
5. Comparer les deux modèles et énoncer qu'elle est, selon vous, le meilleur modèle pour estimer la pureté de notre liquide.

En réalité, le R^2 n'est pas une mesure satisfaisante pour comparer deux modèles, car cela ne tient pas compte du nombre de paramètres présents de ce dernier. On utilise donc souvent le **coefficient de détermination ajusté** R_{aj}^2 pour comparer deux modèles qui emploient un nombre différents de paramètres :

$$R_{aj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1},$$

où n désigne la taille de notre échantillon et p le nombre de variables.

6. Montrer que l'on peut également ré-écrire le coefficient de détermination multiple comme

$$R_{aj}^2 = 1 - \frac{\frac{SCE}{n - (p + 1)}}{\frac{SCT}{n - 1}}.$$

Montrant ainsi que nous avons bien un rapport de deux variances.

Différence significative

Dans cette partie, on cherche à étudier si le modèle quadratique est significativement meilleur que le modèle linéaire. En toute généralité, on cherche à savoir si le modèle à p variables, noté Ω_p , est plus ou moins intéressant qu'un modèle qui contient ces p variables plus une autre, *i.e.* qu'un modèle à $p + 1$ variables, noté Ω_{p+1} . On parle alors de **modèles emboîtés**.

On formule donc le test suivant

H_0 : le modèle Ω_p est valide v.s. le modèle Ω_{p+1} est valide.

La statistique de test F_{test} que l'on considère est définie par

$$F_{\text{test}} = \frac{\frac{SCE(\Omega_p) - SCE(\Omega_{p+1})}{1}}{\frac{SCE(\Omega_{p+1})}{n - (p + 2)}},$$

où $p + 2$ représente le nombre de paramètre du modèle Ω_{p+1} et 1 correspond à la différence, en terme de nombre de paramètres, des modèles Ω_{p+1} et Ω_p .

Cette statistique de test va suivre, sous l'hypothèse H_0 , une loi de Fisher à respectivement 1 et $n - (p + 2)$ degrés de liberté.

1. Effectuer le test et dire si oui ou non le modèle *quadratique* est significativement meilleur que le modèle *linéaire*.

Un autre critère d'évaluation

Il existe d'autres mesures ou critères permettant de comparer des modèles et qui sont plus généraux que le coefficient de détermination. On peut citer le critère **AIC** pour **Akaike Information Criterion** [Akaike, 1974] ou encore le critère **BIC** pour **Bayesian Information Criterion** [Schwarz, 1978].


Ce critère est définie par

$$\text{BIC} = -2 \ln(\hat{\ell}) + (p + 2) \ln(n),$$

où $\hat{\ell}$ désigne le maximum de vraisemblance de nos données. Que cela soit le critère AIC ou BIC, ces deux critères doivent être minimisés afin d'atteindre le meilleur modèle possible.

1. Montrer le critère BIC peut également s'écrire

$$\text{BIC} = n(\ln(2\pi) + 1) + n \ln \left(\frac{SCR}{n} \right) + (p + 2) \ln(n).$$

2. Evaluer et comparer le BIC du modèle linéaire et du modèle quadratique à l'aide de la relation précédente. Est-ce cohérent avec l'observation effectuée avec le critère du R^2 ?
3. Estimer le BIC de vos modèles avec la commande  suivante

```
BIC([nom_du_modèle])
```

En pratique, le critère BIC favorise les modèles qui dépendent de peu de paramètres. Il existe d'autres critères de comparaisons de modèles mais ... nous allons nous arrêter là.

Références

- [Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6) :716–723.
- [Schwarz, 1978] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) :461 – 464.