

Modèles Linéaires

Correction 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

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

Enfin, on rappelle que la "variance" de notre modèle peut être décomposée en la somme de deux termes

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SCT} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SCE} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SCR},$$

où

- **SCT** : somme des carrés **totaux**
- **SCE** : somme des carrés **expliqués** par le modèle
- **SCR** : somme des carrés **résiduels**

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. Rappeler la définition du R^2 .

Le coefficient de détermination est mesuré comme étant la part de la variance expliquée par le modèle par rapport à la variance totale dans les données, *i.e.*,

$$R^2 = \frac{\frac{SCE}{n}}{\frac{SCT}{n}} = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$

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  pour déterminer ces éléments.*

```
# Pour charger un jeu de données
data = read.csv("../data/purity.csv", sep=";", header = TRUE)
colnames(data) = c("Purity", "FilterTime")
n = nrow(data)
# Régression linéaire
mymodel_sim = lm(Purity~FilterTime,data)
summary(mymodel_sim)
```

```
##
## Call:
## lm(formula = Purity ~ FilterTime, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.8984 -4.5298 -0.9656  4.8308 12.6786
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -12.473      3.556  -3.507  0.00252 **
## FilterTime     6.237      0.296  21.072  3.9e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.831 on 18 degrees of freedom
## Multiple R-squared:  0.961, Adjusted R-squared:  0.9589
## F-statistic: 444 on 1 and 18 DF, p-value: 3.904e-14
```

Nous pouvons formuler les remarques suivantes :

- le modèle est globalement significatif comme le montre la p -value de 10^{-3} associée au F -test
- le paramètre associé au temps de filtration est significatif
- le R^2 est égal 0.961, nous avons donc un bon modèle.

3. Faire de même avec le modèle de régression dit quadratique.

```
# On ajoute le carré du temps de filtration à notre jeu de données
data$FilterTime2 = (data$FilterTime)^2
# Régression linéaire quadratique
mymodel_quad = lm(Purity~.,data)
summary(mymodel_quad)

##
## Call:
## lm(formula = Purity ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.6939 -4.0450 -0.7959  3.9422 11.7547
##
## Coefficients:
```

```
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.24631    4.91168  -1.068 0.300396
## FilterTime   4.22809    1.04862   4.032 0.000865 ***
## FilterTime2  0.10095    0.05085   1.985 0.063510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.333 on 17 degrees of freedom
## Multiple R-squared:  0.9684, Adjusted R-squared:  0.9647
## F-statistic: 260.3 on 2 and 17 DF,  p-value: 1.781e-13
```

Nous pouvons formuler les remarques suivantes :

- le modèle est globalement significatif,
- le paramètre associé au temps de filtration est significatif au seuil de 5%,
- le paramètre associé au carré du temps de filtration n'est pas significatif au seuil de 5%, mais il n'est pas loin d'être significatif.
- le R^2 est égal 0.9684, nous avons donc un bon modèle.

4. Les paramètres du modèle sont-ils significatifs ?

Les paramètres du modèles (hors constante du modèle) sont significatifs pour le modèle de régression linéaire simple. Pour le modèle régression linéaire quadratique, le paramètre associé au carré du temps de filtration ne l'est pas.

5. Comparer les deux modèles et énoncer qu'elle est, selon vous, le meilleur modèle pour estimer la pureté de notre liquide.

Si on se base sur le critère du R^2 , le second modèle est préférable au premier.

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

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

On va simplement appliquer la définition du R^2 pour trouver la relation

$$\begin{aligned} R_{aj}^2 &= 1 - (1 - R^2) \frac{n - 1}{n - p - 1}, \\ &= 1 - \left(1 - \frac{SCE}{SCT}\right) \frac{n - 1}{n - p - 1}, \\ &\quad \downarrow \text{ or } \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \\ &= 1 - \frac{SCR}{SCT} \frac{n - 1}{n - p - 1}, \\ R_{aj}^2 &= 1 - \frac{\frac{SCR}{n - (p + 1)}}{\frac{SCT}{n - 1}} \end{aligned}$$

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

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

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*.

On peut légitimement se demander si la différence est significative étant donnée que le modèle quadratique voit son second paramètre peu significatif. Nous allons conduire notre analyse sur  directement.

```
# On commence par extraire les résidus associés à chaque modèle
res_sim = mymodel_sim$residuals
res_quad = mymodel_quad$residuals

# On calcule ensuite la valeur de SCR pour les deux modèles
SCR_sim = sum(res_sim^2)
SCR_quad = sum(res_quad^2)

# On calcule la valeur de la statistique de test
F_test = ((SCR_sim - SCR_quad)/1)/(SCR_quad/(n-3))

# Calcul de la p-valeur
1-pf(F_test,1,n-3)

## [1] 0.06351
```

Ce test ne permet pas de rejeter l'hypothèse H_0 et montre que le modèle quadratique n'est pas significativement meilleur que le modèle linéaire simple. On peut remarquer que la p -valeur de ce test est pile égale à celle du test de significativité du paramètre que l'on a cherché à intégrer. Ce n'est pas un hasard dans ce cas là de modèles emboîtés ! Mais on ne cherchera pas à expliquer pourquoi.

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

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

où \hat{L} 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

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

Etant donné un échantillon $S = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$. On rappelle que la méthode de maximisation de la vraisemblance consiste à trouver les paramètres β et σ^2 optimaux qui maximisent :

$$\ell(S, \beta, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\|\mathbf{y} - \mathbf{X}\beta\|_2^2}{2\sigma^2}.$$

Les paramètres optimaux sont

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad \text{et} \quad \sigma^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2}{n} = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}{n}.$$

Ainsi, le logarithme du maximum de vraisemblance est égale à

$$\begin{aligned} \hat{\ell}(S, \hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\hat{\sigma}^2) - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}{2\hat{\sigma}^2}, \\ &\quad \downarrow \text{ en utilisant les estimateurs} \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \left(\frac{1}{n} \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \right) - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}{\frac{1}{n} \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}, \\ &\quad \downarrow \text{ après simplification} \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \left(\frac{SCR}{n} \right) - \frac{n}{2}, \end{aligned}$$

car $\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SCR$.

Or

$$BIC = -2\ln(\hat{L}) + (p+1)\ln(n),$$

En utilisant l'expression du maximum de vraisemblance, on obtient

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

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 ?

```
# On utilise les valeurs de SCR précédemment calculées.

## Pour le modèle simple
p = 1 # une seule variable
BIC_simple = n*(log(2*pi)+1) + n*log(SCR_sim/n) + (p+2)*log(n)
BIC_simple

## [1] 140.4945

## Pour le modèle quadratique
p = 2 # deux variables
BIC_quad = n*(log(2*pi)+1) + n*log(SCR_quad/n) + (p+2)*log(n)
BIC_quad

## [1] 139.3207
```

Le BIC du modèle quadratique est également légèrement plus faible, ce qui est cohérent avec l'observation faite quant au R^2_{aj} .

3. Estimer le BIC de vos modèles avec la commande  suivante

```
# Pour le modèle linéaire
BIC(mymodel_sim)

## [1] 140.4945

# Pour le modèle quadratique
BIC(mymodel_quad)

## [1] 139.3207
```

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.