Modèles Linéaires

Correction TD 4 : Régression Multiple : Généralités et prédictions
Licence 3 MIASHS

Guillaume Metzler, Francesco Amato, Alejandro Rivera
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr ; francesco.amato@univ-lyon2.fr
alejandro.rivera@univ-lyon2.fr

Résumé

Le but de la présente fiche est de poursuivre notre travail et étude sur la régression linéaire multiple et d'"étudier les prédictions effectuées par le modèle. Plus précisément :

- on cherchera à reproduire les outputs de la fonction *summary* d'un modèle de régression multiple,
- on construira un intervalle de confiance sur les prédictions du modèle.

1 Modèle linéaire multiple

Dans cette section, on considère le modèle multiple suivant

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

où $\mathbf{y} \in \mathbb{R}^n$, $\beta \in \mathbb{R}^{p+1}$, $\mathbf{X} \in^{n\times(p+1)}$ et $\varepsilon \in \mathbb{R}^n$ représentent respectivement, la variable dépendante, le vecteur des paramètres du modèle, matrice de design et le vecteur des erreurs du modèles.

On considère un jeu de données où le but est de prédire le nombre de spectateurs (*attendance*)assistant à un match de baseball en fonctions de diverses variables come la *température*, le nombre de places assises, la taille du stade,... . Le jeu de données utilisé est *attendance.csv* qui est attaché au sujet.

```r
    # Pour charger un jeu de données
    data = read.csv("../data/attendance.csv", header = TRUE)
    n = nrow(data)
    p = ncol(data) - 1
    # Régression linéaire
    mymodel = lm(attendance~.,data)
    summary(mymodel)
```

```
##
## Call:
## lm(formula = attendance ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -974.31 -280.59  -36.55  315.90 1067.98
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1239.40512  327.07553   3.789 0.000202 ***
## temperature     -4.03087    7.76298  -0.519 0.604189
## promotion       47.92031   66.49461   0.721 0.471992
## weekend         32.96119   63.49093   0.519 0.604255
## seats            0.34933    0.04671   7.479  2.6e-12 ***
## size             0.34210    0.03394  10.079  < 2e-16 ***
## rateofwins      -1.80828    2.85917  -0.632 0.527844
## rateofoppwins    4.01839    2.97794   1.349 0.178802
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 434.2 on 192 degrees of freedom
## Multiple R-squared:  0.5545,Adjusted R-squared:  0.5383
## F-statistic: 34.14 on 7 and 192 DF,  p-value: < 2.2e-16
```

1. Commenter les sorties du modèles : dire si ce dernier est globalement significatif ou non. Identifier les variables significatives.

   On peut formuler les remarques suivantes sur les paramètres du modèles :

   - La température a un impact négatif sur le nombre de spectateur, l'augmentation de cette température de une unité, fait perdre 4 spectateurs environ.
   - Lorsque l'on effectue une offre promotionnelle (*promotion = 1*), on gagne

environ 48 spectateurs à ce match par rapport à un match sans offre promotionnelle.
- Même constat pour la variable *weekend* indiquant que les match le week-end attirent environ 32 spectateurs en plus que la semaine.

En revanche on remarque que seules les variables *seats* et *size* sont significatives dans le modèle. Le modèle reste globalement significatif comme le montre la p-valeur associée au test de Fisher. Cependant la qualité global du modèle reste moyenne étant donnés les valeurs des R^2.

2. A l'aide de vos connaissances, reproduire les sorties du logiciel : coefficients, erreurs standards (il s'agit de l'écart-type de l'estimateur), t_{test} et la p-value. On évaluera également l'écart type des valeurs résiduelles, le coefficient de détermination (ajusté) et on effectue le test de Fisher associé à la significativité global du modèle en précisant les paramètres de degrés de liberté qui sont utilisés.

   On effectue directement les calculs sous R. Les définitions des différents objets ont été données en cours.

   La variance de l'estimateur $\text{Var}[\hat{\boldsymbol{\beta}}]$ est donnée par

$$\text{Var}[\hat{\boldsymbol{\beta}}] = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1},$$

   où $\hat{\sigma}^2 = \dfrac{1}{n-p-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$.

```r
# Calcul de la variance de beta_hat

beta_hat = mymodel$coefficients
residuals = mymodel$residuals
sigma2_hat = sum(residuals^2)/(n-p-1)

X = as.matrix(data[,-which(colnames(data)=="attendance")])
X = cbind(c(rep(1,n)),X)
var_beta_hat  = diag(sigma2_hat*solve(t(X)%*%X))

# Pour retrouver l'écart-type
sigma_beta_hat = sqrt(var_beta_hat)
sigma_beta_hat
```

```
##                temperature      promotion        weekend          seats
##  327.07553197   7.76297538   66.49460793   63.49092833    0.04670626
##          size   rateofwins rateofoppwins
##    0.03394251   2.85916668    2.97794169
```

On en déduit directement les valeurs des statistiques de test de student associée aux coefficients β_j par

$$t_{\text{test}}(\hat{\beta}_j) = \frac{\hat{\beta}_j}{\sqrt{\text{Var}[\hat{\beta}_j]}}.$$

```
# Calcul des statistiques de test

t_test = beta_hat / sigma_beta_hat
t_test
```

```
##   (Intercept)   temperature     promotion       weekend         seats
##     3.7893544    -0.5192433     0.7206647     0.5191480     7.4793261
##          size    rateofwins rateofoppwins
##    10.0788964    -0.6324515     1.3493864
```

On en déduit les p-valeurs associées à la statistique de Student

```
# Calcul des p-valeurs

p_values = 2*(1-pt(abs(t_test),n-p-1))
p_values
```

```
##   (Intercept)   temperature     promotion       weekend         seats
##  2.020612e-04  6.041891e-01  4.719924e-01  6.042555e-01  2.601697e-12
##          size    rateofwins rateofoppwins
##  0.000000e+00  5.278444e-01  1.788019e-01
```

On regarde maintenant le coefficient de détermination (ajusté), définie par

$$R^2 = \frac{SCE}{SCT}, \quad \text{et} \quad R^2_{\text{aj}} = 1 - \frac{SCR}{SCT} \times \frac{n-1}{n-p-1}.$$

```
# On commence par calculer les quantités annexes
SCR = sigma2_hat*(n-p-1)
SCT = (n-1)*var(data$attendance)
SCE = SCT - SCR

# Calcul des R_sqare
R_square = SCE/SCT
R_square
```

```
## [1] 0.5545109
```

```r
R_square_adj = 1 - (SCR/SCT)*(n-1)/(n-p-1)
R_square_adj
```

```
## [1] 0.5382691
```

On termine en effectuant le test de Fisher qui permet de tester la significativité globale du modèle. La statistique de Fisher est définie comme rapport de la variance expliquée par le modèle avec la variance résiduelle, *i.e.*,

$$F_{\text{test}} = \frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}}.$$

```r
# Calcul du F_test et de la p-value

F_test = (SCE/SCR)*(n-p-1)/p
F_test
```

```
## [1] 34.141
```

```r
p_value = 1-pf(F_test,p,n-p-1)
p_value
```

```
## [1] 0
```

3. Evaluer le BIC du modèle.

   On rappelle la formule du déterminée au précédent TD

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

```r
# Calcul du BIC
BIC = n*(log(2*pi)+1) + n*log(SCR/n) + (p+2)*log(n)
BIC
```

```
## [1] 3036.522
```

```r
# Comparaison avec la fonction BIC de R
BIC(mymodel)
```

```
## [1] 3036.522
```

4. Exclure toutes les variables significatives du modèle et réapprendre le modèle et évaluer à nouveau son BIC. Est-ce un meilleur modèle ?

   On ne va donc conserver que deux variables pour notre nouveau modèle : *size* et *seats* et reprendre notre analyse.

```
# Regression linéaire
mymodel_sim = lm(attendance~size+seats,data)

n = nrow(data)
p = 2

# Calcul de SCR
SCR_sim = sum(mymodel_sim$residuals^2)

# Calcul du BIC
BIC_sim = n*(log(2*pi)+1) + n*log(SCR_sim/n) + (p+2)*log(n)
BIC_sim
```

```
## [1] 3013.673
```

   Le nouveau modèle a donc un BIC plus faible, il est donc plus intéressant.

Nous verrons plus tard qu'il ne s'agit pas de la façon optimale de rechercher le meilleur modèle que l'on peut construire à partir d'un ensemble de variables (prochain TD).

2 Intervalles de confiance sur la prédiction

Pour cet exemple, on se placera dans le cas du modèle linéaire simple :

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}.$$

Cela nous permettra d'avoir une représentation graphique des résultats, mais la méthodologie à appliquer reste exactement la même dans le cas de la régression multiple.

Pour cela on va considérer le même jeu de données, mais on cherchera uniquement à prédire le nombre de spectateurs (*attendance*) en fonction de la taille du stade (*size*).

```r
    # Pour charger un jeu de données
    data_bis = data[,c("attendance","size")]
    # Régression linéaire
    mymodel = lm(attendance~size,data_bis)
    summary(mymodel)
```

```
##
## Call:
## lm(formula = attendance ~ size, data = data_bis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1307.25  -291.32     3.05   340.36  1120.57
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.390e+03  1.397e+02   9.947   <2e-16 ***
## size        4.264e-01  3.555e-02  11.992   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 487.6 on 198 degrees of freedom
## Multiple R-squared:  0.4207,Adjusted R-squared:  0.4178
## F-statistic: 143.8 on 1 and 198 DF,  p-value: < 2.2e-16
```

On peut vérifier que le modèle appris est bien globalement significatif, l'étude effectuée est donc pertinente.

Intervalle de confiance sur l'espérance de la prédiction

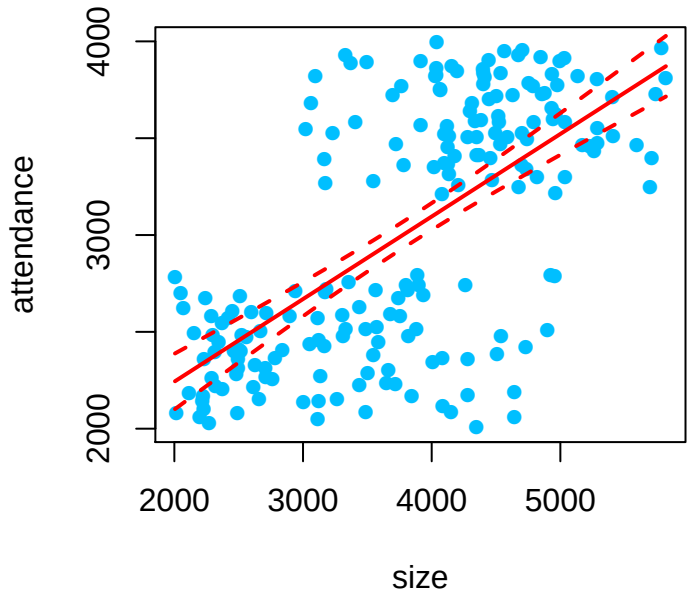
Dans un premier temps, on souhaite construire un intervalle de confiance sur l'espérance de $\mathbb{E}[Y]$ que l'on pourra représenter comme suit.

```r
plot(attendance~size,data=data_bis, pch = 16, col = "deepskyblue")

# On génère échantillon qui nous servira à construire notre IC.

size=seq(min(data_bis$size),max(data_bis$size),length=100)
grille<-data.frame(size)
```

```r
# Prédiction et intervalle de confiance
ICdte<-predict(mymodel,new=grille,interval="confidence",level=0.95)
matlines(grille$size,cbind(ICdte),lty=c(1,2,2), col = "red", lwd =2)
```



On peut montrer que l'intervalle de confiance de niveau $1-\alpha$ sur $\mathbb{E}[y_{\text{new}}]$

$$\hat{y}_{\text{new}} \pm t_{1-\alpha/2,n-p-1}\hat{\sigma}\sqrt{\mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}}}.$$

La démonstration est incluse dans la preuve de la section suivante.

1. Expliquer pourquoi, cette intervalle de confiance n'est pas symétrique. On pourra regarder le terme $\mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$ dans le cas d'un modèle simple pour comprendre le comportement de l'intervalle de confiance.

   Cet intervalle de confiance est centrée autour de la droite de régression mais n'est pas symétrique para rapport à la droite de régression.
   En effet, cela s"explique par le fait que notre production sera plus précise, en moyenne si le nouveau vecteur est proche de la valeur moyenne des observations x_i, mais, on s'en éloigne, la marge d'erreur de notre intervalle de confiance sera plus grande.

Vérifions en faisant les calculs pour le modèle linéaire simple. On se concentre uniquement sur le terme sous la racine.

On a

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \text{donc} \quad \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}.$$

Pour alléger les notations pour la suite, on va poser

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \quad \text{et} \quad \bar{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2.$$

Ainsi, on a

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & n\bar{x^2} \end{pmatrix}$$

Ainsi, l'inverse de cette matrice est donnée par

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{n\,\mathrm{Var}[X]} \begin{pmatrix} \bar{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

On considère maintenant notre vecteur $\mathbf{x}_{\text{new}} = \begin{pmatrix} 1 \\ x_{\text{new}} \end{pmatrix}$ et on calcule la quantité sous la racine, ce qui nous donne :

$$\begin{aligned}
\mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{\text{new}} &= \frac{1}{n\,\mathrm{Var}[X]} \begin{pmatrix} 1 & x_{\text{new}} \end{pmatrix} \begin{pmatrix} \bar{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ x_{\text{new}} \end{pmatrix}, \\
&= \frac{1}{n\,\mathrm{Var}[X]} \begin{pmatrix} 1 & x_{\text{new}} \end{pmatrix} \begin{pmatrix} \bar{x^2} - \bar{x}x_{\text{new}} \\ -\bar{x} + x_{\text{new}} \end{pmatrix}, \\
&= \frac{1}{n\,\mathrm{Var}[X]} \left(\underbrace{\bar{x^2}}_{} - \bar{x}x_{\text{new}} - \bar{x}x_{\text{new}} + x_{\text{new}}^2 \right), \\
&\quad \downarrow \ \mathrm{Var}[X] = \bar{x^2} - \bar{x}^2 \\
&= \frac{1}{n\,\mathrm{Var}[X]} \left(\mathrm{Var}[X] + \bar{x}^2 - 2\bar{x}x_{\text{new}} + x_{\text{new}}^2 \right), \\
&= \frac{1}{n} \left(1 + \frac{(\bar{x} - x_{\text{new}})^2}{\mathrm{Var}[X]} \right).
\end{aligned}$$

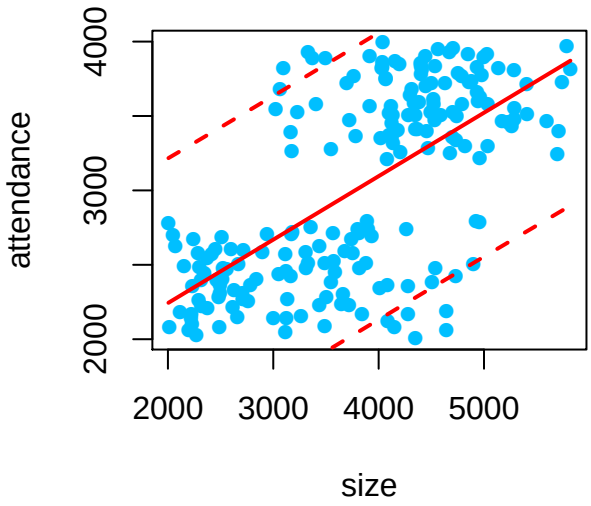
Intervalle de confiance sur la prédiction

On souhaite maintenant construire un intervalle de confiance pour une réponse individuelle Y_i.

```
plot(attendance~size,data=data_bis, pch = 16, col = "deepskyblue")

# On génère échantillon qui nous servira à construire notre IC.
size=seq(min(data_bis$size),max(data_bis$size),length=100)
grille<-data.frame(size)

# Calcul des prédictions ainsi que l'intervalle de confiance.
ICprev<-predict(mymodel,new=grille,interval="pred",level=0.95)
matlines(grille$size,cbind(ICprev),lty=c(1,2,2),col="red",lwd =2)
```



On considère une nouvelle donnée $\mathbf{x}_{\text{new}}$ et on note $\hat{y}_{\text{new}}$ sa prédiction. On sait alors que

$$\hat{y}_{\text{new}} = \hat{\beta}\mathbf{x}_{\text{new}},$$

1. Déterminer l'espérance de $\hat{y}_{\text{new}}$

   On a directement

$$\mathbb{E}[\hat{y}_{\text{new}}] = \mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}.$$

2. Déterminer la variance $\text{Var}[\hat{y}_{\text{new}}]$. La démonstration est similaire à celle consistant à calculer la variance de l'estimateur $\hat{\boldsymbol{\beta}}$. On utilisera le fait que

$$\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

$$\begin{aligned}
\text{Var}[\hat{y}_{\text{new}}] &= \mathbb{E}[(\hat{y}_{\text{new}} - \mathbb{E}[\hat{y}_{\text{new}}])^2], \\
&= \mathbb{E}[(\mathbf{x}_{\text{new}}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2], \\
&= \mathbb{E}[\mathbf{x}_{\text{new}}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbf{x}_{\text{new}}], \\
&= \mathbf{x}_{\text{new}}^\top \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top]\mathbf{x}_{\text{new}}, \\
&= \mathbf{x}_{\text{new}}^\top \text{Var}[\hat{\boldsymbol{\beta}}]\mathbf{x}_{\text{new}}, \\
\text{Var}[\hat{y}_{\text{new}}] &= \sigma^2\mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}}.
\end{aligned}$$

3. En déduire la distribution de $y_{\text{new}} - \hat{y}_{\text{new}}$. On utilisera le fait que les données sont *i.i.d.*.
   On en déduire directement que

$$\hat{y}_{\text{new}} - y_{\text{new}} \sim \mathcal{N}(0, \sigma^2(1 + \mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}})).$$

4. On admettra que la variable aléatoire $T = \dfrac{y_{\text{new}} - \hat{y}_{\text{new}}}{\sqrt{\text{Var}[y_{\text{new}} - \hat{y}_{\text{new}}]}}$ suit une loi de Student à $n - p - 1$ degrés de libertés où p représente le nombre de variables dans le modèle.
   En déduire un intervalle de confiance de niveau $1 - \alpha$ sur la prédiction d'une nouvelle observation.

   L'itervalle de confiance est directement donné par

$$\hat{y}_{\text{new}} \pm t_{1-\alpha/2,n-p-1}\hat{\sigma}\sqrt{1 + 1 + \mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}}}.$$

5. Expliquer pourquoi cet intervalle de confiance est plus grand que le précédent.

   Cet intervalle de confiance est le même que le précédent, on aura simplement ajouté 1 à la valeur $\mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$. Cela montre que l'intervalle de confiance associé à une prédiction individuelle est plus grand (plus d'incertitude) qu'une prédiction en valeur moyenne.