Modèle Linéaire - Régression

Qualité du modèle

Thierry Dhorne

Institut Universitaire de Technologie de Vannes
Université de Bretagne Sud

Année Universitaire 2014-2015

Somme des carrés des écarts totaux

- la variabilité totale de Y est quantifiée par

$$SCET = \sum_{r=1}^{n} (Y_r - \overline{Y})^2$$

- cette variabilité peut être décomposée par l'intermédiaire du modèle

→ pour chaque observation, on décompose l'écart total

$$Y_r - \overline{Y}$$

→ en faisant intervenir $\hat{Y}_r$ le prédicteur de Y_r

- on a alors

$$Y_r - \overline{Y} = (Y_r - \hat{Y}_r) + (\hat{Y}_r - \overline{Y})$$

Décomposition de la SCET

- on a donc

$$
\begin{aligned}
SCET &= \sum_{r=1}^{n} (Y_r - \overline{Y})^2 \\
&= \sum_{r=1}^{n} [(Y_r - \hat{Y}_r) + (\hat{Y}_r - \overline{Y})]^2 \\
&= \sum_{r=1}^{n} [(Y_r - \hat{Y}_r)^2 + (\hat{Y}_r - \overline{Y})^2 + 2(Y_r - \hat{Y}_r)(\hat{Y}_r - \overline{Y})] \\
&= \sum_{r=1}^{n} (Y_r - \hat{Y}_r)^2 + \sum_{r=1}^{n} (\hat{Y}_r - \overline{Y})^2 + 2\sum_{r=1}^{n} (Y_r - \hat{Y}_r)(\hat{Y}_r - \overline{Y})
\end{aligned}
$$

$$
\begin{aligned}
\sum_{r=1}^{n}(Y_r-\hat{Y}_r)(\hat{Y}_r-\overline{Y}) &= \sum_{r=1}^{n}(Y_r-B_0-B_1x_r)(B_0+B_1x_r-\overline{Y}) \\
&= \sum_{r=1}^{n}(Y_r-B_0-B_1x_r)(B_0+B_1x_r-(B_0+B_1\overline{x})) \\
&= \sum_{r=1}^{n}(Y_r-B_0-B_1x_r)\,B_1\,(x_r-\overline{x}) \\
&= B_1\sum_{r=1}^{n}Y_r-\overline{Y}+\overline{Y}-B_0-B_1x_r)(x_r-\overline{x}) \\
&= B_1\sum_{r=1}^{n}(Y_r-\overline{Y}-B_1(\overline{x}-x_r))(x_r-\overline{x}) \\
&= B_1\sum_{r=1}^{n}(Y_r-\overline{Y})(x_r-\overline{x})-B_1^2\sum_{r=1}^{n}(\overline{x}-x_r)^2 \\
&= B_1\sum_{r=1}^{n}(Y_r-\overline{Y})(x_r-\overline{x})-B_1\sum_{r=1}^{n}(\overline{Y}-Y_r)(x_r-\overline{x}) \\
&= 0
\end{aligned}
$$

- la somme des produits s'annule

Équation d'analyse de la variance

- on a donc l'égalité suivante

$$\sum_{r=1}^{n}(Y_r-\overline{Y})^2=\sum_{r=1}^{n}(\hat{Y}_r-\overline{Y})^2+\sum_{r=1}^{n}(Y_r-\hat{Y}_r)^2$$

→ qui s'appelle *équation d'analyse de la variance*

→ qui s'écrit sous forme synthétique

$$SCET = SCEM + SCER$$

- $SCET=\sum_{r=1}^{n}(Y_r-\overline{Y})^2$ est la somme des carrés des écarts totaux
- $SCEM=\sum_{r=1}^{n}(\hat{Y}_r-\overline{Y})^2$ est la somme des carrés des écarts (expliqués par le) modèle
- $SCER=\sum_{r=1}^{n}(Y_r-\hat{Y}_r)^2$ est la somme des carrés des écarts résiduels

- le modèle est il explicatif ?
- → le choix de x comme prédicteur est il justifié ?
- principe
- → $SCEM$ du même ordre que $SCET$
- → $SCER$ le plus petit possible
- critère du R^2

$$R^2 = \frac{SCEM}{SCET}$$

- → si R^2 est proche(?) de 1, le modèle est « intéressant »
- ★ simple mais peu fiable

Validation du modèle

Test de Fisher

- le test de validation de Fisher basé sur la statistique

→ $F = \frac{SCEM/1}{SCER/(n-2)}$

→ le plus grand possible ????

- le F suit (sous H_0) un loi de Fisher $\mathcal{F}_{1,n-2}$

→ significatif (presque toujours ?)

★ sérieux mais peu efficace

```
             Df Sum Sq Mean Sq F value Pr(>F)
ptotal        1  34.32   34.32   207.1 <2e-16 ***
Residuals    57   9.45    0.17
---
Signif. codes :  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```