CHAPITRE VIII

Prédiction et prévision avec la régression linéaire

Des $(x_i, y_i)_i$ observés, on a tiré une estimation $\hat{Y} = \hat{a}_0 + \hat{a}_1 X$
de la relation théorique $Y = a_0 + a_1 X + e$

Pour une nouvelle observation x_h
Comment ***prévoir***[1] y_h ?

Il est clair que la vraie valeur devrait être :

$$y_h = a_0 + a_1 x_h + e_h$$

Mais cela suppose la parfaite connaissance de:

- a_0
- a_1
- et même e_h la valeur de l'erreur théorique !!!

Ne connaissant que $\hat{a}_0, \hat{a}_1$, 3 possibilités s'offrent donc à nous :

1. Utiliser $\hat{a}_0 + \hat{a}_1 X$ ... comme **prédiction**
2. Utiliser $a_0 + a_1 X$ et rappeler que si $\hat{a}_0 + \hat{a}_1 X$ en est un estimateur, il faut lui associer un intervalle de confiance.
3. Utiliser $a_0 + a_1 X + e$ et se rappeler que l'on dispose d'un estimateur de l'erreur théorique.

Revenons sur chacune de ces stratégies.

[1] Prédiction : affirmer ce qui sera Prévoir : : déterminer ce qui pourrait être.

STRATEGIE 1 : Prédiction ou estimation ponctuelle :

On assimile y_h à $\hat{y}_h$ la valeur donnée par la droite de régression linéaire.

Exemple :

- Hier, on a estimé la relation linéaire $\hat{Y} = \hat{a}_0 + \hat{a}_1 X$
- On observe ce matin $x_h = 0{,}3$.
- Le niveau prévu par la droite de régression est :

$$\hat{y}_h = \hat{a}_0 + \hat{a}_1 x_h = \hat{a}_0 + 0{,}3.\hat{a}_1$$

Pb : Cette valeur n'est en fait qu'une ***prédiction*** possible parmi d'autres[2] puisque l'on sait qu'à chaque estimateur $\boxed{\hat{a}_i}$ correspond un intervalle de confiance ...

STRATEGIE 2 : Moyenne des prévisions, intervalle de confiance

On cherche à fixer y_h à la valeur moyenne que l'on doit théoriquement obtenir :

$$a_0 + a_1 x_h = E[y_h]$$

qui est estimé par $\hat{a}_0 + \hat{a}_1 x_h$.
La prise en compte des erreurs d'estimation des coefficients est aisée puisqu'on a les variances estimées et donc $\hat{s}_{\hat{a}_0}$ *et* $\hat{s}_{\hat{a}_1}$.

En fait, les statistiques nous disent que : $E[y_h] \sim N\left(\hat{y}_h, s^2_{E[y_h]}\right)$

Autrement dit :

$$\frac{E[y_h] - \hat{y}_h}{s_{E[y_h]}} \sim N(0{,}1)$$

et qu'un estimateur de la variance est donnée par :

[2] En fait c'est la prédiction moyenne ....

$$\boxed{\hat{\sigma}^2_{E[y_h]} = \hat{\sigma}^2_{\varepsilon}\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

On a donc : $$\frac{E[y_h] - \hat{y}_h}{\hat{\sigma}_{E[y_h]}} \sim t_{n-2}$$

D'où :

$$\boxed{\hat{y}_h - \hat{\sigma}_{E[y_h]} t_{n-2}^{1-\alpha/2} \leq E[y_h] \leq \hat{y}_h + \hat{\sigma}_{E[y_h]} t_{n-2}^{1-\alpha/2}}$$

avec $t_{n-2}^{1-\alpha/2}$ le seuil de confiance.

L'intervalle est d'autant plus grand que :

- la variance estimée augmente ...
- $(x_h - \bar{x})^2$ augmente ...
- que $x_h << \bar{x}$ ou $x_h >> \bar{x}$.

STRATEGIE 3 : Intervalle de prévision

On cherche à déterminer l'ensemble des prévisions "admissibles" en tenant compte de l'erreur théorique de modélisation.

On a théoriquement : $y_h = a_0 + a_1 x_h + \varepsilon_h$

Qui peut aussi s'écrire : $y_h = E[y_h] + \varepsilon_h$

Or, on sait que :

$$E[y_h] \sim N\left(\hat{y}_h, \sigma^2_{E[y_h]}\right)$$

$$\varepsilon \sim N\left(0, \sigma^2_{\varepsilon}\right)$$

Somme de 2 variables normales et indépendantes, on a :

- y_h est une loi normale
- de moyenne la somme des moyennes
- de variance la somme des variances (l'erreur est non corrélée avec la variable exogène)

Conclusion :

$$y_h = E[y_h] + \varepsilon_h \sim N\left(\hat{y}_h, \sigma^2_{E[y_h]} + \sigma^2_{\varepsilon}\right)$$

et donc

$$\frac{y_h - \hat{y}_h}{\sqrt{\sigma^2_{E[y_h]} + \sigma^2_{\varepsilon}}} \sim N(0,1)$$

En remplaçant les variances par leurs estimateurs :

$$\frac{y_h - \hat{y}_h}{\sqrt{\hat{\sigma}^2_{E[y_h]} + \hat{\sigma}^2_{\varepsilon}}} \sim t_{n-2}$$

On obtient donc l'intervalle :

$$\boxed{\hat{y}_h - \sqrt{\hat{\sigma}^2_{E[y_h]} + \hat{\sigma}^2_{\varepsilon}}\, t^{1-\alpha/2}_{n-2} \leq y_h \leq \hat{y}_h + \sqrt{\hat{\sigma}^2_{E[y_h]} + \hat{\sigma}^2_{\varepsilon}}\, t^{1-\alpha/2}_{n-2}}$$

D'un point de vue théorique :

- la stratégie 1 : non satisfaisante,
- la stratégie 2 : si on assume l'erreur théorique,
- la stratégie 3 : l'idéal.