Chapitre 11 : Régression linéaire - Résumé

Afin d'accepter un modèle de régression simple, deux tests sont nécessaires. En premier lieu, un **test de signification (global/Individuel)** est effectué pour mettre en évidence l'utilité de la variable X dans le modèle (son impact sur Y). Par la suite, si le modèle est significatif, on passe à l'étape de **validation** qui permet de connaître les procédures à entreprendre pour améliorer le modèle. Ainsi, un modèle peut être corrigé et réajusté plusieurs fois avant d'obtenir un modèle qui ne peut pas être amélioré davantage.

I-Calcul des estimateurs

Cas particulier : Régression simple

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \text{ et } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$S_{XX} = \sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$$

$$S_{YY} = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - n\bar{X}\bar{Y}$$

II-Test de signification

a-Test global (loi de Fisher)

$H_0: \beta_1 = 0 \ \ vs \ \ H_1: \beta_1 \neq 0.$

L'hypothèse **$H_0 : \beta_1 = 0$** est **rejetée**, au seuil α, si **$F_0 > F_{1,n-2}(\alpha)$**.
Lorsque H_0 est **rejetée**, le modèle est **globalement significatif**, donc la variable X est **significative**.
On peut aussi effectuer un test de signification individuel pour la variable X. C'est un test équivalent au test global pour une régression simple.

Tableau d'analyse de variance – Régression simple

Source de variation	Somme des carrés (khi-2)	Deg liberté	Moyenne des carrés	F (statistique)	p-value
Régression (modèle)	$\boldsymbol{SS_R} = \sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{XY}$	1	$\boldsymbol{MS_R} = \boldsymbol{SS_R}/1$	$F_0 = \boldsymbol{MS_R} / \boldsymbol{MS_E}$	$P(F \geq F_0)$
Résidus (Erreur)	$\boldsymbol{SS_E} = \sum(y_i - \hat{y}_i)^2 = \boldsymbol{S_{YY}} - \boldsymbol{SS_R}$	n-2	$\boldsymbol{MS_E} = \boldsymbol{SS_E}$ /n-2		
Totale	$\boldsymbol{S_{YY}} = \boldsymbol{SC_T} = \sum(y_i - \bar{y})^2$	n-1			

.

b-Tests Individuels (loi de Student)

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$. H_0 est rejetée (X significative) si $|\boldsymbol{T_0}| = \left|\frac{\hat{\beta}_1}{s(\hat{\beta}_1)}\right| > \boldsymbol{t_{n-2}(\alpha/2)}$.

c-Intervalle de confiance pour β_i (i = 0 à 1)

$$\beta_i \in \hat{\beta}_i \mp t_{n-2}(\alpha/2) s(\hat{\beta}_i)$$

$$s(\hat{\beta}_0) = \sqrt{MS_E(\frac{1}{n} + \frac{\bar{X}^2}{S_{SX}})}; \quad s(\hat{\beta}_1) = \sqrt{\frac{MS_E}{S_{XX}}}.$$

àIII-Validation et amélioration du modèle

a- Cœfficient de détermination R^2

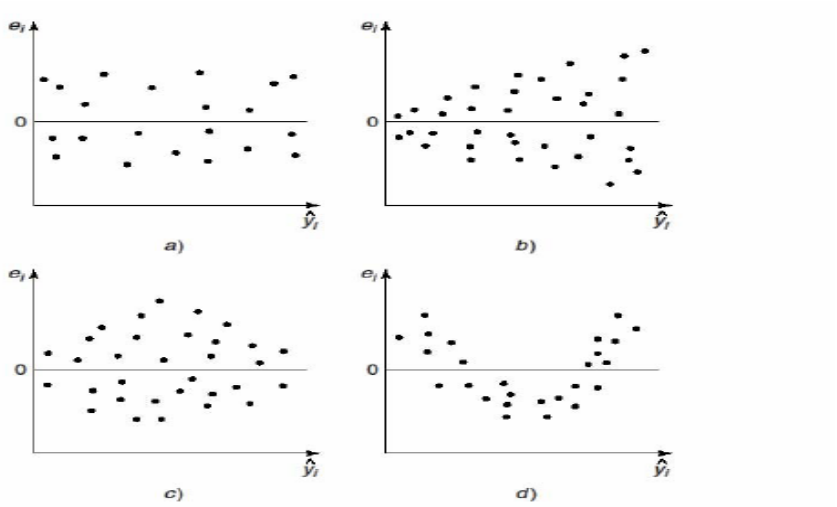
$$R^2 = SS_R / S_{YY} = 1 - SS_E / S_{YY} \qquad R^2_{ajusté} = \frac{(n-1)R^2-1}{n-2}$$

En général, $\mathbf{R^2 > 0.7}$ ➔ le modèle est bon mais pourrait être amélioré (augmenter la valeur de R^2). Pour cela, il faut ajouter de nouvelles variables dans le modèle. Une analyse des résidus donnerait une idée des variables oubliées dans l'ancien modèle et qui pourraient contribuer à son amélioration.

b- Analyse des résidus

Les conditions, sur les résidus, qu'un modèle doit vérifier pour être valide sont entre autres :

- Indépendance des résidus (à l'aide d'un test d'indépendance)
- Normalité des résidus i.e., E suit N(0, σ^2) avec $\hat{\sigma}^2 = MS_E$.
- Absence de tendance particulière. Si, par exemple, les résidus affichent une forme parabolique, ceci veut dire qu'une variable de type X^2 doit être incorporée afin d'améliorer la performance. Voir figure ci-dessous.
- Absence de points aberrants (qui s'éloignent considérablement des autres points). Si de tels points existent, on peut réajuster le modèle en les ignorant. Ainsi plusieurs caractéristiques pourraient être améliorées comme la valeur de R^2 et la normalité des résidus.



a) Pas de tendance particulière, le modèle est satisfaisant, pas d'amélioration à apporter.
b) En entonnoir.
c) A deux arcs.
d) Non linéaire : forme parabolique ➔ introduire X^2.

IV-Intervalle de confiance/prévision :

$$\text{Confiance} : E(Y|X = x_0) \in \hat{y}_0 \mp t_{n-2}(\alpha/2)\sqrt{MS_E(\frac{1}{n} + \frac{(\bar{X}-x_0)^2}{S_{XX}})}$$

$$\text{Prévison} : (Y|X = x_0) \in \hat{y}_0 \mp t_{n-2}(\alpha/2)\sqrt{MS_E(1 + \frac{1}{n} + \frac{(\bar{X}-x_0)^2}{S_{XX}})}$$

x0-x

V-Corrélation échantillonnale :

$$r_{X,Y} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \text{ donc } (r_{X,Y})^2 = R^2$$