Chapitre 03 : Diagnostic et validations des modèles

1\ Contrôle de l'aptitude du Modèle :

1-1\ Analyse des Résidus :

Dans un modèle de régression linéaire le résidu ε_i est la différence entre la valeur réel et la valeur ajustée

$$\varepsilon_i = y_i - \hat{y}_i .$$

- l'analyse des résidus permet de vérifier la validation des hypothèses du modèle de régression linéaire.

* Les hypothèses du modèle linéaire :

- Linéarité : la relation entre x et y doit être linéaire.
- Indépendance des résidus : aucune autocorrélation.
- Normalité des résidus : Les erreurs doivent suivre une loi normale.
- Homos-cédasticité : La variance des erreurs doit être constante.

1-2\ Graphiques de résidus :

L'examen de la validité des hypothèses du modèle, se fait à partir du graphe des résidus ε.

Scanné avec CamScanner

1\ Linéarité de la relation :

On peut juger et analyser la linéarité entre X et Y en visualisant le graphique des couples (x_i, y_i) (nuage des points)

2\ Normalité des résidus :

Nous pouvons vérifier l'hypothèse de normalité en utilisant des méthodes graphiques comme :

a - Graphique Q-Q plot (quantile-quantile plot)

Est appelé "Droite de Henry" et un graphique "nuage des points" qui vise à confronter les quantils de la distribution empirique et les quantiles d'une distribution théorique normale, de moyenne et l'écart type estimés sur les valeurs observées.

Si la distribution est compatible avec la loi normale, les points forment une droite.

b\ Histogramme des résidus :

On obtient également une représentation des résidus du modèle, dont on peut vérifier la compatibilité avec la distribution Gaussienne.

- on peut aussi vérifier l'hypothèse de normalité par des tests comme :

c\ Test de Jarque-Bera :

Le test d'hypothèse s'écrit de la manière suivante :

$$\begin{cases} H_0 : \varepsilon \text{ suit une loi normale} \\ H_1 : \varepsilon \text{ ne suit pas une loi normale.} \end{cases}$$

Scanné avec CamScanner

La statistique de Jarque-Bera est : $T = \frac{n-p-1}{6}\left(g_1^2 + \frac{g_2^2}{4}\right)$

avec la RC au risque α s'écrit :

$$T > \chi^2_{1-\alpha}(2) \quad \text{sous } H_0.$$

d) Test de Shapiro-Wilk :

il basée sur la statistique : $W = \frac{\left[\sum_{i=1}^{[\frac{n}{2}]} a_i (x_{n-i+1} - x_i)\right]^2}{\sum_i (x_i - \bar{x})^2}$

- La RC, rejet de la normalité s'écrit (RC : $W < W_{crit}$)

e) Test de Kolmogorov-Smirnov :

3) Homoscédasticité :

- Homoscédasticité : la variance résiduelle S^2_ε est constante
- Hétéroscédasticité : la variance résiduelle S^2_ε n'est pas constante

- Nous proposons plusieurs graphiques possibles pour détecter une hétéroscédasticité, il est recommandée de tracer les résidus studentisés t_i^* en fonction des $\hat{y}_i$.

si une structure apparaît (tendance, cône, vagues), l'hypothèse d'homoscédasticité risque fort de ne pas être vérifiée. ou bien par l'examen de la série chronologique tracée, après avoir divisé les résidus en deux parties et calculé la variance de chaque partie, si les deux variance sont égale alors l'hypothèse d'homoscédasticité est vérifiée.

Scanné avec CamScanner

4\ Indépendance :

La détection de l'autocorrélation des résidus peut s'effectuer visuellement à l'aide du graphique des résidus comme les graphique "acf", on ne constate aucune structure particulière, et peu de bâtons dépassent les bornes limites, on admet l'indépendance des résidus.

ou bien par le graphique des résidus en fonction des numéros d'observations t, par ailleurs, nous pouvons aussi utiliser des tests comme :

a\ Test de Durbin - Watson :

test spécifique à une forme de l'erreur puissant pour cette forme $\varepsilon_i = \rho \varepsilon_{i-1} + v_i$ avec $v_i \rightsquigarrow \mathcal{N}(0, \sigma_v)$

- le test d'hypothèse s'écrit :

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

- on utilise la statistique de Durbin - Watson :

$$d = \frac{\sum_{i=1}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

b\ Test des séquences : " Test de Wald Wolfowitz "

1-3\ Transformation des données :

La transformation des données en régression linéaire est une étape très importante quand les relations entre les variables ne sont pas linéaire ou que certaines hypothèses de la régression linéaire (comme l'homoscédasticité ou la normalité des résidus) ne sont pas respectées.

1-3-1\ Pourquoi transformer les données :

La transformation des données peut aider à rendre :

- la relation entre les variables plus linéaire.
- stabiliser la variance
- Rendre les résidus plus normaux.

1-3-2\ Types de transformation plus courantes :

a\ Logarithmique :

Réduit l'asymétrie positive (biais à droite) et stabilise la variance, utile pour des données avec de grandes valeurs.

b\ Racine carrée :

Pour des données biaisées à droite, moins extrême que la "log".

c\ Carré :

Pour des données biaisées à gauche.

d\ Inverse :

Surtout pour RLM, rapproche les valeurs extrêmes de la moyenne, utile pour des données avec

Scanné avec CamScanner

des valeurs très grandes ou petits.

* Méthodes avancées :

. Box - Cox :

Famille de transformation (incluant le log, carré...) qui trouve la puissance optimale pour rendre les données normales.

. Yeo - Johnson :

Extension de Box - Cox qui gère aussi les valeurs négatives.

. Standardisation : (Z - score)

Centrer les données à une moyenne de 0 et une variance de 1 (échelles différents), utile surtout pour RLM.

1-3-3) Étapes pratiques :

- Visualiser les données avec un nuage des points.
- Identifier la forme de la relation (linéaire, exponentielle,..
- Choisir une transformation appropriée pour X, y ou les deux.
- Appliquer la RL sur les données transformées.
- Vérifier les hypothèses des résidus.
- Interpréter les coefficients dans le contexte transformé

Exemple :

@ - Supposons que on mesurée la croissance y en fonction du temps X et que les points montrant une croissance exponentielle.

Scanné avec CamScanner

- Relation non linéaire.

$y = 2e^{0,3x}$

→ transformation logarithmique : $\log(y) = \log 2 + 0,3x$

② - pour $y > 0$, tester y^λ avec différents λ :

. $\lambda = 0 \Rightarrow \log(\lambda) \Rightarrow$ transformation log.

. $\lambda = \frac{1}{2} \Rightarrow \sqrt{y} \Rightarrow$ " racine carrée.

. $\lambda = 1 \Rightarrow$ pas de transformation.

③ si : $y = \frac{1}{x}$, alors transformer x en $\frac{1}{x}$ peut linéariser la relation.

Scanné avec CamScanner