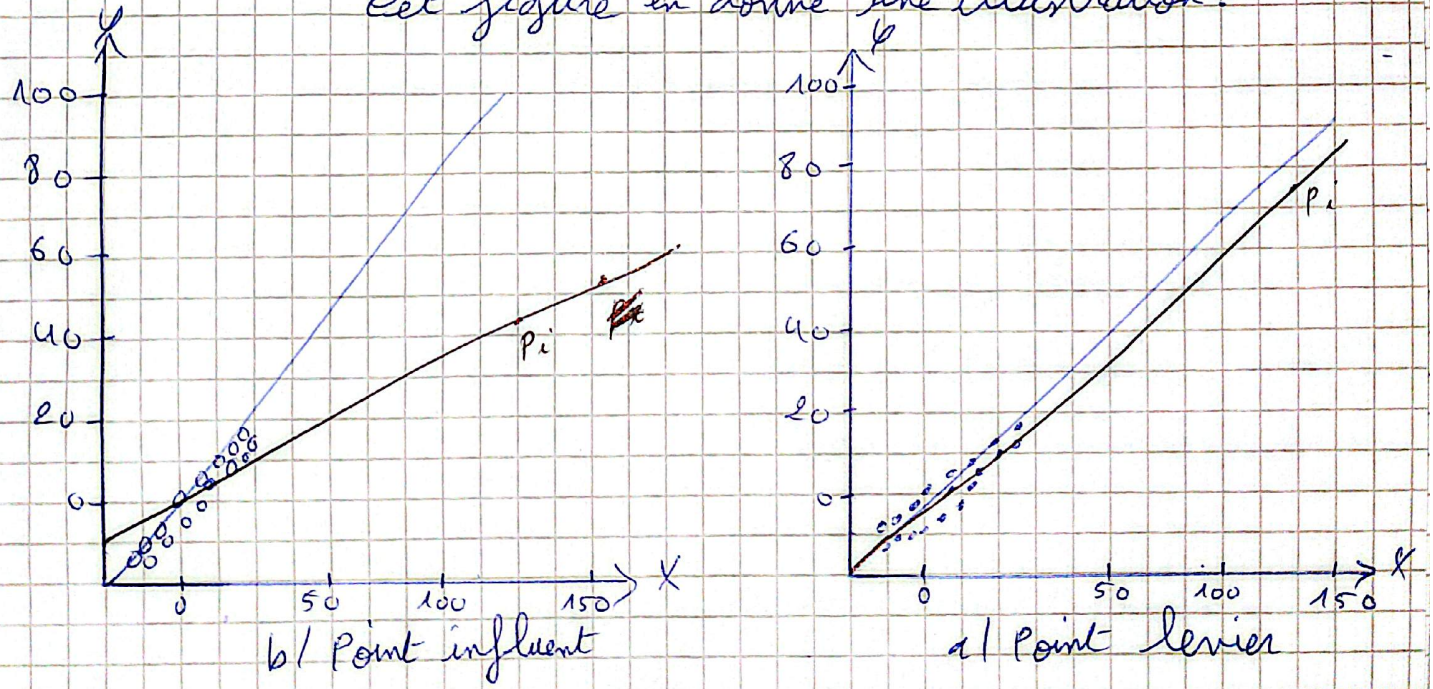
Chapitre 04 : Diagnostic de régression et Points d'influence.

1/ Points de levier et Points d'influents :

Il s'agit ici d'indentifier les points qui exercent une influence significative sur l'équation de la droite de régression, on entend par cela que l'équation de la droite de régression change de façon importante lorsque l'on supprime ces points.

Cet figure en donne une illustration :



- Point Levier :

On appelle point levier un point dont la coordonnée sur l'axe X est significativement différente de celles des autres points.

Scanné avec CamScanner

- Règle pour identifier les points leviers :

on considère qu'un point i est un point levier si :

$$h_{ii} > \frac{4}{n} \quad \text{avec} : h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$$

h_{ii} : appèle le score de levier.

- Points d'influence :

Un point d'influence est une observation qui, si elle était retirée de l'analyse, changerait de manière significative les coefficients de régression.

- La distance de Cook :

Est également une statistique utilisée pour évaluer l'influence d'un point sur la droite de régression. Cette distance est donnée par la formule suivante :

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_{j(i)} - \hat{y}_j)^2}{2 s^2}$$

on peut écrire : $D_i = \frac{r_i^2}{2} \times \frac{h_{ii}}{1 - h_{ii}}$

- on considère qu'une point i est un point d'influence si : $D_i > \frac{4}{n}$ (ou $D_i > 1$).

- Remarque :

La différence majeure entre un point influent et un point levier réside dans le fait que dans le cas du point influent le résidu de celui-ci est atypique (significativement plus grand en valeur absolue) que celui des autres points, alors que dans le cas d'un point levier, son résidus n'est pas atypique par rapport aux résidus des autres points.

Scanné avec CamScanner

2\ - Méthode de Box - Cox :

La transformation de Box - Cox est une méthode statistique permettant de stabiliser la variance et de rendre les données plus proches d'une distribution normale. Elle a été proposée par Box et Cox en 1964.

- Définition :

pour des données positive $y_1, y_2, y_3, \dots, y_n$ ($y_i > 0$), la transformation Box - Cox est définie par :

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & , \lambda \neq 0 \\ \log(y_i) & , \lambda = 0 \end{cases}$$

où :

- λ est le paramètre de transformation à estimer par MV
- si les données contiennent des zéros ou des valeurs négatives, il faut ajouter une constant ($c + y$).

2 - 2 - 1\ Choix de λ :

- Pour $\lambda = 1$: Pas de transformation (données origina
- Pour $\lambda = 0$: transformation logarithmèque
- Pour $\lambda = -1$: transformation inverse

Scanné avec CamScanner

2-2-2\ Etapes pour appliquer Box - Cox :

1- Vérifier que toutes les valeurs sont positives.
2- Définir le modèle statistique ($y = f(x)$).
3- Estimer λ par MVE.
4- Transformer les données avec λ choisi.
5- Vérifier la distribution après transformation par :
   - histogramme
   - Q-Q-plot.
6- Appliquer l'analyse statistique sur les données transformées.

3\ Régression linéaire Généralisée (GLM) :

La régression généralisée est une extension de la régression linéaire classique, qui permet de modéliser des variables dépendantes non normales et de relier la moyenne de la variable réponse à une combinaison linéaire des variables explicatives.

- La GLM permet de modéliser correctement ces cas :
  . Y est binaire (0/1) => exemple : réussite / échec.
  . Y est un compte => exemple : nombre de visites incidents.
  . Y est une proportion => exemple : pourcentage de succès.

3-1\ Structure d'une GLM :

supposons une variable réponse Y de famille exponentielle linéaire, ainsi la fonction de densité s'exprime comme :

$$f(y) = C(y, \phi) \exp\left[\frac{y\theta - a(\theta)}{\phi}\right]$$

Scanné avec CamScanner

avec : $g(\mu) = X^T \beta$, appelé la fonction de lien,

qu'est une transformation de la moyenne μ. La fonction de lien $g(\mu)$ est quand à elle reliée linéairement avec les variables explicative exprimées dans le vecteur X.

3-2\ Fonction de lien canonique :

si $g(\mu) = \theta$, alors g est appelée le lien canonique correspondant à $a(\theta)$. Dans une telle situation $\theta = X^T \beta$.

Exemple :

Trouvez le lien canonique $g(\mu)$ pour une loi de Poisson

- Nous savons que : $P[Y = y] = \frac{\lambda^y e^{-\lambda}}{y!}$

et que : $c(y, \phi) = -\ln(y!)$

$\theta = \ln(\lambda)$

$a(\theta) = \lambda = \exp\theta$

$\phi = 1$

Ainsi, puisque la moyenne $\mu = \lambda$ pour une loi de Poisson : $\theta = \ln(\lambda) = \ln(\mu)$ et donc $g(\mu) = \ln(\lambda)$

on appelle ce lien, le lien logarithmique.

3-3\ La structure linéaire (Prediction) :

$M = X\beta = \beta_0 + \beta_1 X_1 + \cdots + X_p \beta_p$

où : X : matrice des prédicteurs

β : coefficients à estimées.

Scanné avec CamScanner

3-4/ Estimation des paramètres :

Les coefficients β sont estimées par MVE, la fonction de log vraisemblance s'exprime comme :

$$\text{Log } \mathcal{L}(\beta, \phi) = l(\beta, \phi) = \sum_{i=1}^{n} \ln(f_{y_i}, \beta, \phi)$$

$$= \sum_{i=1}^{n} \left(\ln(c(y_i, \phi)) + \frac{y_i \theta_i - a(\theta_i)}{\phi} \right)$$

$$= \frac{1}{\phi} \sum_i [y_i \theta_i - a(\theta_i)] + \sum_i \ln(c(y_i, \phi))$$

avec : $\eta_i = X_i^T \beta = g(\mu_i)$ pour des variables y_i, $i = \overline{1,n}$ indépendantes.

- En utilisent la règle de dérivation en chaîne, pour estimer le MLE des β_j ; $j = \overline{1,p}$.

Scanné avec CamScanner