Rappel: Modèle de regression linéaire Simple et Multiple

Base: The social production of Criminal Homicide
13 variables ↓ 20 observations ↓ 20 cities

Les VI: City / Pop / Homic / Poor % pauvreté
ville Am nb de pers en milliers victime/100 000 hab

Obj: expliquer pourquoi le taux d'Homicide est élevé dans certaines villes que dans d'autres

. même nombre d'homicide mais nb pop ≠ ⇒ Cau l'une tient compte de la criminalité

Le taux d'homicide standardisé, rationalisé (normalisé) par rapport à la population, a permis de comparer des villes de population différente.

Histo de la distribution d'Homicide: montre que la variable endogène Homicide ne suit pas une dist normale. inf/sup/=/

Estimation Kernel densité/densité du noyau: Approche enracinée dans la méthodologie de l'histogramme: Estime la fonction de densité en un point x en utilisant les obs voisi

Soit la fonction de poids: $K(x) = \begin{cases} 1/2 & \text{si } |x|<1 \\ 0 & \text{sinon} \end{cases}$
de noyau de Kernel

d'estimateur densité Kernel: $f(x) = \frac{1}{n.h} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$
fonction additif de Kernel / noyau ←

Certains fonctions de Kernel:

Uniform: $\frac{1}{2}\mathbb{1}(|x|<1)$ Epanechnikov: $\frac{3}{4}(1-x^2)\mathbb{1}(|x|\leq 1)$

Gaussian: $\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$ Biweight: $\frac{15}{16}(1-x^2)\mathbb{1}(|x|\leq 1)$

La mon normalité: valeur élevée de min et max

Asymétrie de la dist: par les indicateurs moy et Mediane (par la moyenne de centile alternative)

Mediane = Moyenne: dist sym Skewness = 0

Moyenne > Mediane: dist affectée par des val. ext. (asym ≥ 0) ≥ 0 Skew > 0

Moyenne < Mediane: dist affecté par des outliers (asym ≤ 0) <0 Skew < 0

Test de normalité: JB- test: S: Skewness EK: Excès de Kurtosis

Ho: $S = EK = 0$ où $EK = \text{Kurtosis} - 3 \Leftrightarrow \text{Kurtosis} = 3$

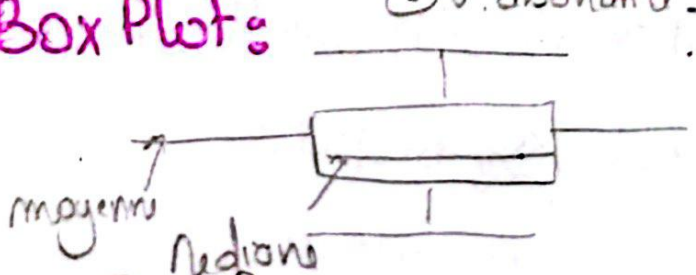
$\Leftrightarrow$ dist normale

La stat: $JB = N\left(\frac{\hat{S}^2}{6} + \frac{\hat{EK}^2}{24}\right) \sim \chi^2(2)$

avec: $\hat{S}^2 = \frac{\frac{1}{N}\sum(y_i-\bar{y})^3}{\left(\frac{1}{N}\sum(y_i-\bar{y})^2\right)^{3/2}}$ $\hat{EK} = \frac{\frac{1}{N}\sum(y_i-\bar{y})^4}{\left(\frac{1}{N}\sum(y_i-\bar{y})^2\right)^2} - 3$

→ p-value < α% $\Rightarrow$ distribution non normale

Box Plot:



⊙ V. aberrante → à cause de son taux d'homicide exceptionnellement élevé dans l'échantillon → a un très grand résidus

$\Rightarrow$ ε de la regression ne suivent pas normalement distribuées

$\Rightarrow$ On ne peut pas appliquer les distributions tctf

regression homic poor : homic = 0,9438 Poor − 0,8152 ← exogène

B0 ↗ 1%
↳ hom ↗ 0,9

Car pas de ville sans pauvreté ← Pas de pauv / { Le taux dans les villes 0% pauvreté ↵

RMS = MES : l'estimation de la variance d'erreur

$$RMS = S^2 = \frac{\sum \hat{e}_i^2}{n-p} \ (= 29,53)$$

Root MSE : déviation standard : mesure de diagnostic de la regression

$$Root MSE = \sqrt{S^2} \ (= 5,43)$$ Erreur stand de la regression

RootMSE élevée ⇒ résultats de regression non fiables

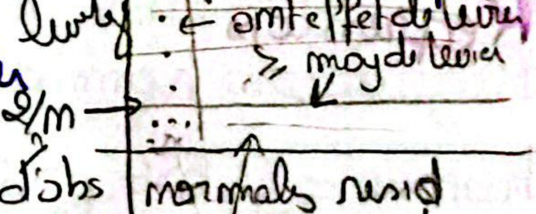
homichat : valeur prédite   res : sont les résidus de l'estim

⇒ Pas de corrélation entre les val. prédites et résidus

La distance de Cook, effet levier / leverage, les résidus standarisés : outils qui détectent les valeurs aberrantes

- Effet de levier : l'éloignement d'un point % à $\bar{Y}$

Le levier $h_i \geq 2 . \frac{p+1}{n}$: règle de détectation : valeur de levier à partir de laquelle on commence à s'inquiéter



Horiz : les valeurs moy d'effet de levier

Vertical : les valeurs moy des résidus normalisés

v. aberrante → n'a pas trop de levier

- Distance de Cook : influence global des y_i sur les $\hat{y}_i$
se calcul à l'aide des valeurs de levier et résidus standarisés

Comparer D à la f-stat à (n-p) ddl

<20% : peu d'inf 50% : grande inf

région critique : $Di > \frac{4}{m-p-1}$

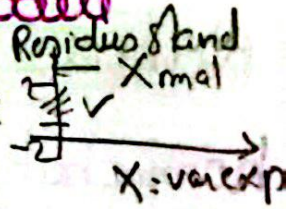
- Résidus Stamdardisés : Compar ya $\hat{y}$

$ri = \frac{ei}{S\sqrt{1-hii}}$ $\forall i=1,..,m$ $hii = \frac{1}{m} + \frac{(xi-\bar{x})^2}{\sum (xi-\bar{x})^2}$ [2.2]

résidus stamd ~ les résidus normalisés sauf $V(ri)=1$

Résidu élevé ⇒ point mal reconstruit par le modèle

Pour détecter les valeurs atypiques : on trace le nuage



- Résidus Studentisés : $es_{(i)} = \frac{ei}{S_{(i)}\sqrt{1-hii}}$ $\forall i=1,..,m$

Les permettent de mieux évaluer l'importance d'un résidu ei pour détecter les val atypiques [2.2] ✓

RStudent élevé ⇒ point mal modélisé et RStamd faible

point exagérément influe ⇒ attire les résultats vers lui

Regression homic pour pop :

homic = 0,656 pour + 0,022 pop − 3,899

La pauvreté exerce un effet sur homic plus fort que pop

pauv-pour en % et population - poop en milliers : On calcule les beta pour rendre comparable la force des coeff

beta : les coeff après une 1ère conversion Z-scores

↳ Les coeff bêta sont directement comparables

Reg homic poor pop, beta

homic = 0,656 poor + 0,022 pop − 3,899 | Les β sont inversés pop est plus colinéaire avec la var homic

$\beta = 0,35$ $\beta = 0,76$

$\tilde{\beta}_K = \hat{\beta}_K \frac{s_{x_K}}{s_y}$ $K = 1, \dots, K-1$

$\in [-1,1]$

$\tilde{\beta}_K$: coeff standardisé de la Kième var exp

$\hat{\beta}_K$: coeff OLS habituel

s : déviation standard

Comparaison des modèles simples et multiple :

coeff de la var. exogène poor : 0,94 ⟶ 0,65

↓ hautement biaisé car la var omise pop est ≥ 0 et fortement corrélée à pop

F-test = 36,3 et p-value = 0 ⇒ Rejette l'hypo de base $^{Coef=0}$

⇒ Les 2 paramètres sont jointement significatifs ≠ 0 dans la population

↓ tout régresseur est apte à prédire la variable endog homic

$y_i = \alpha + \varepsilon_i \quad \forall i = 1, \dots, n$

↳ bruit imprévisible

Les étapes pour le calcul du F-test :

Estimer le modèle (OLS) → SCR_{mc} = 135,276 → RSS

Estimer le modèle avec seulement l'intercept → SCR_c = 712,94

Reg homic : homic = 6,9 / Estimation OLS de l'intercept $= \bar{y}$

Stat du test : $\hat{F} = \frac{SCR_c - SCR_{mc}}{SCR_{mc}} \times \frac{N-K}{q} \sim f(q, N-K)$

N : nb Obs

K : nbr des paramètres

nbr des param restreints $q = K-1$

On peut tester la signification des 2 paramètres par une autre approche:

Calcul des résidus OLS restreints: $\hat{\varepsilon}_i = y_i - \hat{a}$ définis sous l'hypothèse $H_0: \beta_1 = \beta_2 = 0$, X_{i1} et X_{i2} doivent être non corrélés

Regression par OLS $\hat{\varepsilon}_i$ sur X_{i1} et X_{i2}

Calcul LM-stat = $NR^2 \sim \chi^2(q)$ q: nbr de contraintes à tester

$R^2 \uparrow \Rightarrow \hat{\varepsilon}_i$ est signif corrélé avec X_{i1} et X_{i2} $\Rightarrow$ $NR^2 > \chi^2(q)$ / $p_v < 0,05$ $\Rightarrow$ rejet H_0 $\Rightarrow$ significative

il faut inclure X_{i1} et X_{i2} dans la regression

RMSE non proche de 0 $\Rightarrow$ le modèle évalué n'est pas meilleur en terme d'exactitude