INF232 :
ANALYSE DE DONNEES

REGRESSION LINÉAIRE SIMPLE

Cas d’erreur gaussiennes

Plan

- Introduction
- Estimateur du maximum de vraisemblance
- Lois usuelles
- Lois des estimateurs
- Intervalles de confiance
- Validation du modèle de régression linéaire
- Prévision
- Cas pratique

Introduction

Mieux que les expressions des estimateurs et des variances pour la régression linéaire, la connaissance de lois permet de retrouver leur région de confiance et ainsi effectuer des tests d'hypothèses. Le but ici est de pouvoir généraliser les estimations faites sur notre échantillon à la population et ainsi voir si notre modèle de régression linéaire peut être extrapolé ou généralisé à la population.

Tout d'abord faisons les hypothèses suivantes :

$$\begin{cases} (H1): \varepsilon_i \rightsquigarrow N(0,1) \\ (H2): \varepsilon_i \ mutuellement\ indépendantes \end{cases}$$

Le modèle de régression linéaire simple devient un modèle paramétrique à paramètre $(\beta_1, \beta_2, \sigma^2)$. La loi de ε_i étant connue, les lois des y_i deviennent $y_i \rightsquigarrow N(\beta_1 + \beta_2 * x_i, \sigma^2)$ et les y_i sont mutuellement indépendantes.

Estimateur du maximum de vraisemblance (1/2)

Pour avoir $(\beta_1, \beta_2, \sigma^2)$ nous allons utiliser le maximum de vraisemblance et chercher les paramètres ($\widehat{\beta_1}$, $\widehat{\beta_2}$, $\widehat{\sigma^2}$) qui maximise notre vraisemblance.

L'expression de la vraisemblance est :

$$\begin{aligned} \mathcal{L}(\beta_1, \beta_2, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}S(\beta_1, \beta_2)\right] \end{aligned}$$

Etant donné que manipuler une telle fonction semble complexe, nous calculons la log vraisemblance : $\log \mathcal{L}(\beta_1, \beta_2, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}S(\beta_1, \beta_2).$

Estimateur du maximum de vraisemblance (1/2)

Nous voulons maximiser cette quantité pour ($\beta_1, \beta_2, \sigma^2$). Les paramètres β_1 et β_2 appartiennent uniquement au terme -S(β_1, β_2) qu'il faut minimiser (car est -S(β_1, β_2) ≤ 0). Or on a déjà les quantités $\widehat{\beta_1}$ et $\widehat{\beta_2}$ des MCO qui minimisent ce terme. On cherche maintenant à maximiser log L ($\beta_1, \beta_2, \sigma^2$) par rapport à σ^2.

$$\frac{\partial \log \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} S(\hat{\beta}_1, \hat{\beta}_2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - |\hat{\beta}_2 x_i)^2$$

On en déduit l'estimateur du maximum de vraisemblance de σ^2par : $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon_i}^2$

Cet estimateur est biaisé car E ($\hat{\sigma}^2$) $= \frac{n-2}{n} \sigma^2$. On doit le rendre sans biais. Ainsi, $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{\varepsilon_i}^2$

Mais pour un nombre d'observation suffisamment grand, ce biais est négligeable.

Lois usuelles

Les lois d’usage constant en régression linéaire sont : **loi de khi-2, loi de Student et loi de Fisher.**

Dans cette partie, nous allons à tour de rôle décrire ces différentes lois et donner un cas d’utilisation pour chacune d’elles.

Lois usuelles : khi-2

Soit x_1, x_2, …, x_n des variables aléatoires iid (indépendantes et identiquement distribuées) telles que $x_i \rightsquigarrow N(0,1)$ $\forall i = 1, 2 \ldots n$.

La loi de $X=\sum_{i=1}^{n} {x_i}^2 \rightsquigarrow \chi^2(n)$. Où $\chi^2(n)$ est la loi du χ^2 à n degré de liberté (ddl).

Cette loi est généralement utilisée pour tester l'indépendance de deux séries statistiques à valeur qualitative (exemple savoir si la région d'origine a une influence sur le statut social. Ici on essaie de mesurer la distance entre la valeur observée et la valeur théorique et on compare le résultat à la valeur théorique du khi-2 à

(c-1)×(l-1) ddl où c et l sont respectivement le nombre de catégories dans la variable région et dans la variables statut social).

Lois usuelles : Student

Soit Z une variable aléatoire telle que Z $\rightsquigarrow N(0,1)$, soit X une variable aléatoire suivant le χ^2(n). La variable aléatoire T = $\frac{Z}{\sqrt{\frac{X}{n}}}$ est appelé loi de Student à n ddl et est noté T $\rightsquigarrow T_n$.

NB : Elles sont au nombre de trois :

- Lorsque n=1, T est une loi de Cauchy et n'a ni espérance, ni variance ;
- Lorsque n=2, T est centrée mais de variance infinie ;
- Pour n $\geq$ 3, T est centrée et de variance $\frac{n}{n-2}$;
- Lorsque n devient grand, par la loi des grands nombres T $\rightsquigarrow N(0,1)$ et $\frac{n}{n-2} = 1$

La loi de Student est généralement utilisée lorsque l'on veut faire un test sur la moyenne. C'est-à-dire tester si la moyenne de la population prend une valeur particulière (Par exemple pour répondre à l'affirmation : en moyenne les étudiants de l'Université de Yaoundé 1 sont plus âgés que les étudiants de Université de Dschang. Il s'agira ici de comparer à travers un test d'hypothèse si les moyennes des deux populations sont égales ou si celle de UY1 est supérieure à celle de Uds. On calculera une statistique de test qui suivra une loi de **Student**. Puis de comparer la valeur calculée à la valeur théorique se trouvant dans la table de loi du **Student** pour le ddl correspondant).

Lois usuelles : Fisher

Soient $U_1 \rightsquigarrow \chi^2(n_1)$ et $U_2 \rightsquigarrow \chi^2(n_2)$ deux variables aléatoires indépendantes. La loi de $F = \frac{U_1/n_1}{U_2/n_2}$ est la loi de Fisher à (n_1, n_2) ddl. On note $F \rightsquigarrow F_{n_2}^{n_1}$.

NB : Pour $n_2 \geq 2$, l'espérance de $F_{n_2}^{n_1}$ est $\frac{n_2}{n_2-2}$. Dans la suite, n2 sera grand de sorte qu'à nouveau la loi des grands nombres implique que $\frac{U_2}{n_2} \rightsquigarrow 1$. Dans ce cas, F peut être vu comme un χ^2 normalisé par son ddl : $F \rightsquigarrow \chi^2_{n1}/n_1$.

L'utilisation de la loi de Fisher se retrouve dans la comparaison des variances. C'est-à-dire tester si les variances de deux populations sont égales (Par exemple pour répondre à l'affirmation : la dispersion des âges des étudiants de l'Université de Yaoundé 1 est supérieure à celle des étudiants de Université de Dschang. Il s'agira ici de comparer à travers un test d'hypothèse si les variances des deux populations sont égales ou si celle de UY1 est supérieure à celle de Uds. On calculera une statistique de test qui suivra une loi de **Fisher**. Puis de comparer la valeur calculée à la valeur théorique se trouvant dans la table de loi du **Fisher** pour le ddl correspondant).

Lois usuelles : Cas d'utilisation dans la régression linéaire

Dans le cadre de la régression linéaire, ces trois lois sont utilisées pour pouvoir tester si notre modèle peut être généralisé à la population avec :

- Un test de Student sur paramètre β_2 (Ici on test si $\beta_2 \neq 0$ avec pour hypothèse nulle $\beta_2 = 0$). Cela permet de voir si effectivement la variable explicative X explique réellement la variable expliqué Y.
- Un test de Fisher qui permet de voir tout comme le test de Student si $\beta_2 \neq 0$. Cela se fait par le calcul d'une statistique de test F (qui sera présenté plus bas) que l'on comparera à la valeur théorique de F_{n-2}^{1}. Ici on utilise implicitement χ^2 pour ce qui est de la variance des erreurs du modèle.

Lois des estimateurs

Notons C, σ_1^2, σ_2^2 sont mes variances et covariance des estimateurs MCO. Et $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$ sont les variances de $\widehat{\beta_1}$ et $\widehat{\beta_2}$.

Les lois des estimateurs sont :

i. $\widehat{\beta} = \begin{pmatrix} \widehat{\beta_1} \\ \widehat{\beta_2} \end{pmatrix} \rightsquigarrow N(\beta, \sigma^2 \times V)$ Où $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ et $V = \frac{1}{\sum(x_i - \bar{x})^2} \times \begin{pmatrix} \sum x_i^2 / n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} = \frac{1}{\sigma^2} \times \begin{pmatrix} \sigma_1^2 & C \\ C & \sigma_2^2 \end{pmatrix}$

ii. $\frac{(n-2)}{\sigma^2}\hat{\sigma}^2 \rightsquigarrow \chi^2_{\text{n-2}}$

iii. $\widehat{\beta}$ et $\hat{\sigma}^2$ sont indépendants.

On aura les lois suivantes :

i. $\frac{\widehat{\beta_1} - \beta_1}{\hat{\sigma}_1^2} \rightsquigarrow T_{\text{n-2}}$

ii. $\frac{\widehat{\beta_2} - \beta_2}{\hat{\sigma}_2^2} \rightsquigarrow T_{\text{n-2}}$

iii. $\frac{1}{2\hat{\sigma}^2}(\hat{\beta} - \beta)' V^{-1} (\hat{\beta} - \beta) \rightsquigarrow F_{n-2}^2$

Cette propriété nous permet de donner l'intervalle de confiance (IC) des estimateurs.

Intervalles de confiance

- Ci-dessus nous avons fait une estimation ponctuelle des paramètres ($\beta_1, \beta_2, \sigma^2$). Or il est plus pratique de donner une plage de valeur dans laquelle la valeur réelle du paramètre pourrait s'y trouver. Cette plage de valeur est l'intervalle de confiance (IC). Ainsi pour nos paramètres ($\beta_1, \beta_2, \sigma^2$) l'estimation par intervalle de confiance est :

i. IC (β_1) = $\widehat{\beta_1} \pm t_{n-2}(1-\frac{\alpha}{2}) \times \hat{\sigma}_1$ Où $t_{n-2}(1-\frac{\alpha}{2})$ est le quantile de niveau $(1-\frac{\alpha}{2})$ de la loi de Student à n-2 ddl

ii. IC (β_2) = $\widehat{\beta_2} \pm t_{n-2}(1-\frac{\alpha}{2}) \times \hat{\sigma}_2$ Où $t_{n-2}(1-\frac{\alpha}{2})$ est le quantile de niveau $(1-\frac{\alpha}{2})$ de la loi de Student à n-2 ddl

iii. IC (σ^2) = $[\frac{(n-2)\widehat{\sigma^2}}{C_{n-2}(1-\frac{\alpha}{2})}; \frac{(n-2)\widehat{\sigma^2}}{C_{n-2}(\frac{\alpha}{2})}]$ Où $C_{n-2}(\frac{\alpha}{2})$ est le quantile de niveau $\frac{\alpha}{2}$ de $\chi^2_{\text{n-2}}$.

Validation du modèle de régression linéaire

Les paramètres de notre modèle étant estimés, il est intéressant de savoir si notre modèle estimé peut-être généralisé ou extrapolé à la population. Pour ce faire, nous avons le choix entre un test de Student de significativité de $\widehat{\beta_2}$ et un test de Fisher de significativité de $\widehat{\beta_2}$ utilisant la somme des carrés des estimés (SCE) et la somme des carrés résiduels (SCR).

Hypothèse de test : $\begin{cases}(H0):\ \beta_2 = 0\\(H1):\ \beta_2 \neq 0\end{cases}$

Statistique de test est : $t = \frac{\widehat{\beta_2}}{\widehat{\sigma}_2} \rightsquigarrow T_{n-2}$

Resultat : Notre modèle sera valide si et seulement si la valeur calculée de la statistique de test en valeur absolue est supérieure à $t_{n-2}(1-\frac{\alpha}{2})$.

Hypothèse de test : $\begin{cases}(H0):\ \beta_2 = 0\\(H1):\ \beta_2 \neq 0\end{cases}$

Statistique de test est : $F = \frac{SCE}{SCR/_{n-2}} \rightsquigarrow F^1_{n-2}$

Resultat : Ensuite nous comparons cette valeur calculée F au quantile de 1 et n-2 ddl à 1-α de confiance. Le modèle sera valide lorsque la valeur calculée est supérieure à la valeur théorique se trouvant dans la table de loi de Fisher.

Prévision

Le but principal de la modélisation étant de faire une représentation simplifiée la réalité d'une part et de faire de la prévision d'autres part, notre modèle de régression linéaire validé doit nous permettre de prédire une nouvelle valeur $\hat{y}_{n+1}$ pour un x_{n+1} donné.

$\hat{y}_{n+1}$ Étant linéaire en $\widehat{\beta_1}$, $\widehat{\beta_2}$ et ε_{n+1}, on peut préciser sa loi : $y_{n+1} - \hat{y}_{n+1} \sim \mathcal{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)\right)$

On ne connait pas σ^2. On l'estime avec $\widehat{\sigma^2}$ et on peut alors donner l'intervalle de confiance de y_{n+1} en utilisant la loi :

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum(x_i - \bar{x})^2}}} \sim \mathcal{T}_{n-2}$$

Et on obtient :

$$\left[\hat{y}_{n+1} \pm t_{n-2}(1 - \alpha/2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\right]$$

Cas pratique : Avec les deux stagiaires

Analyse de la variance :

	ddl	Somme des Carrés	Moyenne des Carrés	F Calculé	F Théorique
Expliqué	1	1,898	1,899	0,104	4,96
Résidus	10	183,019	18,302		
Total	11	184,917			

Prédiction des paramètres du modèle :

	Estimation	Ecart Type ($\hat{\sigma}_i$)	t-test	t-théorique	IC (Borne Inf.)	IC (Borne Sup.)
$\boldsymbol{\beta_1}$	11,99	3,131	3,829847	2,228	5,015	18,966
$\boldsymbol{\beta_2}$	0,108	0,335	0,322011	2,228	-0,639	0,855

Conclusion :

Les valeurs calculées de F et t étant inférieures à leurs valeurs théoriques respectives, nous pouvons affirmer avec une confiance de 95% (ou un risque d'erreur de 5%) que le paramètre $\beta_2 = 0$. Et donc que la note à l'épreuve « A » n'explique pas linéairement la note à l'épreuve « B ». Cela s'observe aussi avec l'appartenance de zéro à l'intervalle de confiance de β_2. Par ailleurs on note qu'en absence de note à l'épreuve « A », le stagiaire aurait eu un peu moins de 12 à l'épreuve « B ».

Cas pratique : Sans les deux stagiaires

Analyse de la variance :

	ddl	Somme des Carrés	Moyenne des Carrés	F Calculé	F Théorique
Expliqué	1	80,677	80,677	33,402	5,32
Résidus	8	19,323	2,415		
Total	9	100,000			

Prédiction des paramètres du modèle :

	Estimation	Ecart Type ($\hat{\sigma}_i$)	t-test	t-théorique	IC (Borne Inf.)	IC (Borne Sup.)
β_1	5,470	1,392	3,928	2,306	2,259	8,681
β_2	0,896	0,155	5,779	2,306	0,539	1,254

Conclusion :

Les valeurs calculées de F et t étant supérieures à leurs valeurs théoriques respectives, nous pouvons affirmer avec une confiance de 95% (ou un risque d'erreur de 5%) que le paramètre $\beta_2 \neq 0$. Et donc que la note à l'épreuve « A » explique linéairement la note à l'épreuve « B ». Et donc pour une augmentation d'une unité de la note à l'épreuve « A », on observera une augmentation de la note à l'épreuve « B » de 0,896. En absence de note à l'épreuve « A » le stagiaire aurait eu un peut plus de 5/20 à l'épreuve « B ».