Eléments de statistiques pour les data sciences

Cours 8: modèle gaussien - régression linéaire - théorème central limite

Dr. Ph. Müllhaupt

IGM - EPFL

plan

modèle gaussien

hypothèses de base

- on considère n variables aléatoires indépendantes

$$Y_1, Y_2, \ldots, Y_n$$

- qui sont supposées distribuées de manière normale (gaussiennes) avec des moyennes différentes μ_i, $i = 1, \ldots, n$, mais des variances identiques σ^2

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \qquad i = 1 \ldots, n$$

- ainsi les différences observées sont hypothétisées provenir des différences de centrage (moyenne) mais pas par l'étalement (variance constante)

modèle gaussien

hypothèses de base

définition

erreur de modèle :

$$\epsilon_i \triangleq Y_i - \mu_i \sim \mathcal{N}(0, \sigma^2)$$

que l'on peut réécrire sous la forme

$$Y_i \triangleq \mu_i + \epsilon_i \qquad \text{avec } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

modéle gaussien

interprétation de la variance

remarque (rappel d'un résultat) : Lorsque deux variables sont indépendantes et identiquement distribuées

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \qquad Y_i' \sim \mathcal{N}(\mu_i, \sigma^2)$$

alors on a le résultat (cf. en utilisant la convolution et les fonctions génératrices)

$$Y_i - Y_i' \sim \mathcal{N}(0, 2\sigma^2)$$

On peut calculer la probabilité suivante (interprétation répétition fréquente de l'expérience)

$$P\{|Y_i - Y_i'| \geq \sigma\} = P\{|Z| > \frac{1}{\sqrt{2}}\} = 0.48$$

Interprétation de la variance lors d'une répétition de mesure : Grosso modo 1 mesure sur 2, on aura un écart supérieur à σ.

modèle gaussien

hypothèse concernant les moyennes μ_i, $i = 1, \ldots, n$

- supposons n mesures
- il y a $n + 1$ paramètres μ_1, μ_2, ..., μ_n et σ
- $\Rightarrow$ il faut faire l'hypothèse d'une relation entre les paramètres pour pouvoir estimer

- moyennes identiques : $\mu_1 = \mu_2 = \ldots = \mu_n = \alpha$: modèle à un échantillon
- deux groupes : $\mu_1 = \mu_2 = \ldots \mu_p = \alpha$ et $\mu_{p+1} = \ldots = \mu_n = \alpha + \beta$: modèle à deux échantillons
- une droite inconnue relie les moyennes $\mu_i = \alpha + \beta\, x_i$, $i = 1, \ldots n$ où x_i sont des constantes : régression linéaire

Ce sont des modèles dits linéaires car les paramètres sont reliés entre eux par une expression linéaire (contrainte linéaire) dans chacun des cas.

modèle gaussien

méthode du maximum de vraisemblance

- trait commun entre les modèles linéaires : utilisation du maximum de vraisemblance
- $\Rightarrow$ déduction de la fonction de coût quadratique J

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\begin{aligned} L(\mu_1, \ldots, \mu_n, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} \\ l(\mu_1, \ldots, \mu_n, \sigma) = \log L &= -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 \\ J &\triangleq \sum_{i=1}^{n} (y_i - \mu_i)^2 = \sum_{i=1}^{n} \epsilon_i \end{aligned}$$

modèle gaussien

méthode du maximum de vraisemblance

Méthode pour obtenir les MLE $\hat{\alpha}$ et $\hat{\beta}$ (estimation de la paramétrisation des moyennes $\mu_i = \mu_i(\alpha, \beta)$, $i = 1, \ldots, n$) :

- J ne dépend pas de σ : minimisation de J
- les μ_i sont fonctions linéaires de α et β
- $\Rightarrow$ système d'équations linéaires à résoudre :

$$\frac{\partial J}{\partial \alpha} = 0 \qquad \frac{\partial J}{\partial \beta} = 0$$

modèle gaussien

méthode du maximum de vraisemblance

définition : valeurs ajustées $\hat{\mu}_i$ et résiduels $\hat{\epsilon}_i$

Une fois les paramètre estimés $\hat{\alpha}$ et $\hat{\beta}$ par la résolution de $\frac{\partial J}{\partial \alpha} = 0$ et $\frac{\partial J}{\partial \beta}$, à cause de la paramétrisation des moyennes μ_i, celles-ci prennent des valeurs bien particulières $\hat{\mu}_i$, $i = 1, \ldots, n$ appelées valeurs ajustées. La quantité

$$\boxed{\hat{\epsilon}_i \triangleq y_i - \hat{\mu}_i}$$

est appelée la $i$ème valeur ajustée.

modèle gaussien

information fournie par les résiduels $\hat{\epsilon}_i$

les résiduels $\hat{\epsilon}_1$, $\hat{\epsilon}_2$, ..., $\hat{\epsilon}_n$ informent sur :

- l'adéquation du modèle probabiliste : utilisation du $\chi^2_{(n-d)}$
- la valeur de la variance σ^2 : variance d'échantillon s^2

modèle gaussien

adéquation statistique

adéquation statistique

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 \sim \chi^2_{(n-q)}$$

- q est le nombre de contraintes
- le nombre de contraintes est égal au nombre de paramètres
- $\Rightarrow$ q est donc le nombre de paramètres α, β, ... à estimer
- L.4 Pearson, sl. 20 : $X \sim \mathcal{N}(0,1)$ et $Z = X^2 \Rightarrow Z \sim \chi^2_{(1)}$
- $X_i \sim \chi^2_{(1)}$, $i = 1, \ldots n \Rightarrow \sum X_i \sim \chi^2_{(n)}$ (cf. exo avec les fct. génératrices).

modèle gaussien

variance d'échantillon

définition : variance d'échantillon

$$s^2 \triangleq \frac{1}{n-q} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{\text{numérateur du } \chi^2}{\text{degrés de liberté}} \tag{1}$$

modèle gaussien

variance d'échantillon

Définissons $V \triangleq (n-q)\frac{s^2}{\sigma^2} \sim \chi^2_{(n-q)}$. La densité de probabilité de V est

$$f(v) = k_\nu \, v^{(\nu/2)-1} e^{-v/2} \qquad \nu = n-q \qquad k_\nu \text{ une constante}$$

changement de variables pour obtenir la densité de probabilité (marginale) de $\sum \hat{\epsilon}_i^2$

$$\begin{aligned} f(v) \cdot \left| \frac{dv}{d\sum \hat{\epsilon}_i^2} \right| &= k_\nu v^{(\nu/2)-1} e^{-v/2} \cdot \frac{1}{\sigma^2} \\ &= k_\nu \left[\frac{\nu s^2}{\sigma^2} \right]^{(\nu/2)-1} \exp\{-\frac{\nu s^2}{2\sigma^2}\} \cdot \frac{1}{\sigma^2} \end{aligned}$$

modèle gaussien

variance d'échantillon

$$l(\sigma) = -\nu \log \sigma - \frac{\nu s^2}{2\sigma^2} \qquad \sigma > 0$$

$l'(\sigma) = 0$ donne

$$s = \sigma$$

commentaire

Si on avait maximiser le logarithme de la vraisemblance sans passer par la distribution marginale des $\hat{\epsilon}_i^2$ on aurait obtenu simplement la variance $\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2$. On aurait pas tenu compte des contraintes $n - q$. En minimisant le logarithme de la vraisemblance de la distribution marginale des $\hat{\epsilon}_i^2$ on obtient la variance d'échantillon qui tient compte des contraintes :

$$\boxed{s^2 \triangleq \frac{1}{n-q} \sum_{i=1}^{n} \hat{\epsilon}_i^2}$$

modèle gaussien

modèle linéaire et inférence sur les paramètres, variance σ^2 connue

les paramètres estimés $\hat{\alpha}$, $\hat{\beta}$, ... sont fonctions linéaires de valeurs observées y_1, ..., y_n :

$$\begin{aligned} \hat{\alpha} &= a_1\, y_1 + a_2\, y_2 + \ldots + a_n\, y_n \\ \hat{\beta} &= b_1\, y_1 + b_2\, y_3 + \ldots + b_n\, y_n \end{aligned}$$

conséquence : $\hat{\alpha}$, $\hat{\beta}$ sont gaussiennes

$$\text{var}(\hat{\alpha}) = \left(\sum a_i^2\right) \text{var}(Y_i) = \sum a_i^2\, \sigma^2$$

$$\boxed{Z \triangleq \frac{\hat{\alpha} - \alpha}{\sqrt{\sigma^2 \sum a_i^2}} \sim \mathcal{N}(0,1)}$$

test de signification + intervalles de confiance

modèle gaussien

modèle linéaire et inférence sur les paramètres, variance σ^2 inconnue

Lorsque la variance n'est pas connue, et que celle-ci est estimée à partir de la variance d'échantillon

$$s^2 = \frac{1}{n-q}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-q}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

il faut utiliser la distribution de Student :

$$\boxed{T \triangleq \frac{\hat{\alpha} - \alpha}{\sqrt{s^2 \sum a_i^2}} \sim t_{(n-q)}}$$

test de signification + intervalles de confiance

modèle à échantillon unique

$\mu_1 = \ldots = \mu_n = \alpha$

- $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$
- $\alpha = \mu_1 = \mu_2 = \ldots \mu_n$
- MLE : $\Rightarrow J = \sum (y_i - \mu_i)^2 = \sum (y_i - \alpha)^2$ à rendre minimum

ceci conduit à

$$\left.\frac{dJ}{d\alpha}\right|_{\alpha=\hat{\alpha}} = -2\sum(y_i - \hat{\alpha}) = -2\sum y_i - 2n\hat{\alpha} = 0$$

$$\hat{\alpha} = \frac{\sum y_i}{n} \triangleq \bar{y}$$

Le MLE du paramètre $\hat{\alpha}$ devient dans ce cas la moyenne des valeurs observées.

$$\hat{\alpha} = \sum a_i y_i \Rightarrow a_i = \frac{1}{n} \qquad \hat{\mu}_1 = \ldots = \hat{\mu}_n = \hat{\alpha} \qquad \boxed{\sum \hat{\epsilon}_i^2 = \sum (y_i - \bar{y})^2}$$

modèle à échantillon unique

$\mu_1 = \ldots = \mu_n = \alpha$

- $\sum a_i^2 = n \left(\frac{1}{n}\right)^2 = \frac{1}{n}$
- $q = 1$

modèle à un échantillon

exemple

Un médicament donne un augementation de pression cardiaque de 22. On a testé un nouveau médicament qui donne sur 10 individus les résultats suivants :

18	27	23	15	18	15	18	20	17	8

$$\bar{y} = \frac{1}{10}(18 + 27 + 23 + 15 + 18 + 15 + 18 + 20 + 17 + 9) = \frac{179}{10} = 17.9$$

La variance n'est pas connue et on l'estime avec la variance d'échantillon avec $d = 1$

$$s^2 = \frac{1}{9}(\sum y_i^2 - \bar{y} \sum y_i) = \frac{1}{9}(3433 - 17.9 \cdot 179) = 25.43$$

$$T = \frac{\hat{\alpha} - 22}{\sqrt{s^2/n}} \sim t_{(n-1)} \quad t = \frac{17.9 - 22}{\sqrt{25.43/10}} = -2.57 \sim t_{(9)}$$

modèle à un échantillon

exemple

	probabilité qu'une valeur supérieure soit atteinte, sans compter le signe								
ν	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
9	0.7027	0.8834	1.383	1.833	2.262	2.685	3.25	3.69	4.781

$$NS = P\{|t_{(9)}| \geq 2.57\} \approx 0.05 - \frac{2.57 - 2.262}{2.685 - 2.262} 0.025 \approx 0.03$$

Il y a de l'évidence que l'effet du nouveau médicament diffère du médicament initial.

modèle à deux échantillons

$\mu_1 = \ldots \mu_p = \alpha,\ \mu_{p+1} = \ldots \mu_n = \alpha + \beta$

- $Y_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2) \qquad j = 1, 2, \ldots, n_i, i = 1, 2$
-

$$\begin{aligned} \mu_{11} &= \mu_{12} = \ldots = \mu_{1\,n_1} = \alpha \\ \mu_{21} &= \mu_{22} = \ldots = \mu_{2\,n_2} = \alpha + \beta \end{aligned}$$

$$\begin{aligned} J &= \sum\sum (y_{ij} - \mu_{ij})^2 \\ &= \sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2 \end{aligned}$$

modèle à deux échantillons

$\partial J/\partial \alpha = 0$, et $\partial J/\partial \beta = 0$

$$\begin{aligned} \hat{\mu}_1 &= \bar{y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j} \\ \hat{\mu}_2 &= \bar{y}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} y_{2j} \end{aligned}$$

$$\hat{\alpha} = \hat{\mu}_1 \qquad \hat{\beta} = \hat{\mu}_2 - \hat{\mu}_1 = \bar{y}_2 - \bar{y}_1$$

valeurs ajustées et résiduels :

$$\begin{aligned} \hat{\mu}_{ij} &= \hat{\mu}_i = \bar{y}_i \\ \hat{\epsilon}_{ij} &= y_{ij} - \hat{\mu}_{ij} = y_{ij} - \bar{y}_i \end{aligned}$$

modèle à deux échantillons

la somme des carrés des résiduels :

$$\begin{aligned}\sum\sum \hat{\epsilon}_{ij}^2 &= \sum\sum (y_{ij} - \bar{y}_i)^2 \\ &= \sum (y_{1j} - \bar{y}_1)^2 + \sum (y_{2j} - \bar{y}_2)^2 \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\end{aligned}$$

avec

$$s_1^2 \triangleq \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2$$

$$s_2^2 \triangleq \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2$$

il y a $q = 2$ paramètres et la formule (1) indique :

$$s^2 = \frac{1}{n-2} \sum \sum \hat{\epsilon}_{ij}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

modèle à deux échantillons

inférence sur $\beta = \mu_2 - \mu_1$

$$\bar{Y}_1 \sim \mathcal{N}\left(\mu_1, \frac{1}{n_1}\sigma^2\right) \qquad \bar{Y}_2 \sim \mathcal{N}\left(\mu_2, \frac{1}{n_2}\sigma^2\right)$$

$\bar{Y}_1$ et $\bar{Y}_2$ sont indépendants $\Rightarrow$

$$\bar{Y}_2 - \bar{Y}_1 \sim \mathcal{N}\left(\mu_2 - \mu_1, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

$$\boxed{T \triangleq \frac{\hat{\beta} - \beta}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n-2)}}$$

modèle à deux échantillons

exemple de O. H. Latter dans *Biometrika* 1902, p. 164

Les coucous pondent des oeufs dans les nids d'autres oiseaux. On compare les nids de rousseroles et les nids de troglodytes.

oeux nids rousseroles			oeux nids troglodytes				
21.2	21.6	21.9	19.8	20.0	20.3	20.8	20.9
22.0	22.0	22.2	20.9	21.0	21.0	21.0	21.2
22.8	22.9	23.2	21.5	22.0	22.0	22.1	22.3

modèle à deux échantillons

exemple

$n_1 = 9\,;\ \bar{y}_1 = 22.2$
$s_1^2 = 0.4225$ (8 d.l.)

$n_2 = 15\,;\ \bar{y}_2 = 21.12$
$s_2^2 = 0.5689$ (14 d.l.)

modèle à deux échantillons

exemple

Inférence pour β

$$T = \frac{\hat{\beta} - \beta}{\sqrt{s^2\left(\frac{1}{9} + \frac{1}{15}\right)}} \sim t_{(n-2)}$$

k avec $\hat{\beta} = \bar{y}_2 - \bar{y}_1 = -1.08$, $n = 24$,

$$s^2 = \frac{8s_1^2 + 14s_2^2}{8 + 14} = 0.5156\,(22\text{d.l.})$$

$$t = \frac{\hat{\beta} - 0}{\sqrt{s^2(1/9 + 1/15)}} = -3.57$$

$$NS = P\{|t_{(22)}| \geq 3.57\} \approx 0.002$$

Si μ_1 était égal à μ_2, une différence aussi grande que celle observée arriverait très rarement. Il y a donc une grande évidence que $\mu_1 \neq \mu_2$.

régression linéaire

le modèle de la droite

- $Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \qquad i = 1, 2, \ldots, n$
- $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$

$$J = \sum (y_i - \mu_i)^2 = \sum (y_i - \alpha - \beta x_i)^2$$

$$\frac{\partial J}{\partial \alpha} = -2 \sum (y_i - \alpha - \beta x_i) \qquad \frac{\partial J}{\partial \beta} = -2 \sum x_i (y_i - \alpha - \beta x_i)$$

$\frac{\partial J}{\partial \alpha} = 0$ donne

$$\sum y_i - n \hat{\alpha} - \hat{\beta} \sum x_i = 0$$

régression linéaire

le modèle de la droite

$\frac{\partial J}{\partial \beta} = 0$ donne

$$\begin{aligned} 0 &= \sum x_i(y_i - \hat{\alpha} - \hat{\beta} x_i) \\ &= \sum x_i(y_i - \bar{y} + \hat{\beta}\,\bar{x} - \hat{\beta} x_i) \\ &= \sum x_i(y_i - \bar{y}) - \hat{\beta} \sum x_i\,(x_i - \bar{x}) \end{aligned}$$

ce qui donne

$$\hat{\beta} = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})} \triangleq \frac{\sigma_{xy}}{\sigma_{xx}}$$

régression linéaire

modèle de la droite — calcul des sommes de produits

$$
\begin{aligned}
\sigma_{xy} &= \sum (y_i - \bar{y}) x_i = \sum (x_i - \bar{x}) y_i = \sum (x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum x_i y_i - n \bar{x} \bar{y} \\
&= \sum x_i y_i - \frac{1}{n} (\sum x_i)(\sum y_i) \\
\sigma_{xx} &= \sum (x_i - \bar{x}) x_i = \sum (x_i - \bar{x})^2 \\
&= \sum x_i^2 - n \bar{x}^2 = \sum x_i^2 - \frac{1}{n} (\sum x_i)^2
\end{aligned}
$$

régression linéaire

estimation de la variance

les valeurs ajustées et les résiduels :

$$\begin{aligned} \hat{\mu}_i &= \hat{\alpha} + \hat{\beta} x_i = \bar{y} + \hat{\beta}(x_i - \bar{x}) \\ \hat{\epsilon}_i &= y_i - \hat{\mu}_i = (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}) \end{aligned}$$

la somme des carrés des résiduels :

$$\begin{aligned} \sum \hat{\epsilon}_i &= [(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})]^2 \\ &= \sum (y_i - \bar{y})^2 - 2\hat{\beta} \sum (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}^2 \sum (x_i - \bar{x})^2 \\ &= \sigma_{yy} - 2\hat{\beta}\sigma_{xy} + \hat{\beta}^2 \sigma_{xx} \end{aligned}$$

comme

$$\hat{\beta} = \frac{\sigma_{xy}}{\sigma_{xx}} \Rightarrow \sum \hat{\epsilon}_i^2 = \sigma_{yy} - \hat{\beta}\sigma_{xy}$$

et

$$s^2 = \frac{1}{n-2} \sum \hat{\epsilon}_i^2$$

régression linéaire

exemple

Soit la table suivante qui représente des valeurs de la quantité y en fonction de l'âge x de la personne.

x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

régression linéaire

exemple

$$\sum x_i = 628 \qquad \sum y_i = 1684$$
$$\sum x_i^2 = 34416 \qquad \sum y_i^2 = 238822 \qquad \sum x_i y_i = 89894$$

$$\bar{x} = 52.33 \qquad \bar{y} = 140.33$$
$$\sigma_{xx} = 1550.67 \qquad \sigma_{yy} = 2500.67 \qquad \sigma_{xy} = 1764.67$$

$$\hat{\beta} = \frac{\sigma_{xy}}{\sigma_{xx}} = 1.138 \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = 80.78$$

régression linéaire

exemple

la droite de régression est donc :

$$\boxed{\hat{y} = 80.78 + 1.138\, x}$$

somme des carrés des résiduels :

$$\sum \hat{\epsilon}_i^2 = \sigma_{yy} - \hat{\beta}\, \sigma_{xy} = 492.47$$

estimée de la variance autour de la droite de régression :

$$s^2 = \frac{1}{n-2} \sum \hat{\epsilon}_i^2 = 49.247 \qquad\qquad n - 2 = 10 \text{ d.l.}$$

régression linéaire

exemple

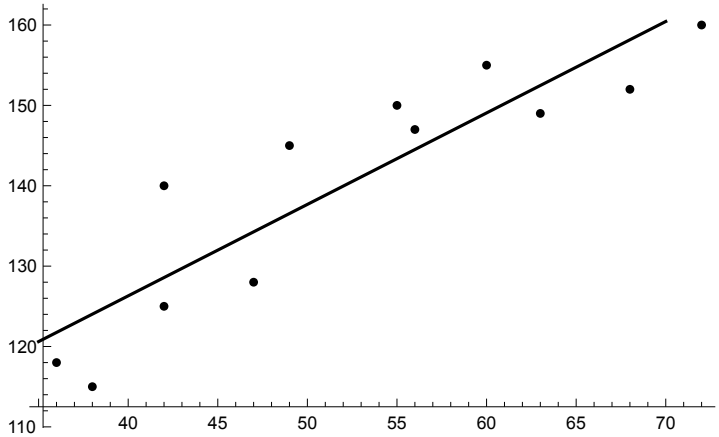

Figure – graphique de la droite de régression

régression linéaire

inférence sur $\hat{\beta}$

$$\hat{\beta} = \frac{\sigma_{xy}}{\sigma_{xx}} = \frac{\sum (x_i - \bar{x}) y_i}{\sigma_{xx}} = \sum a_i y_i$$

avec a_i les constantes

$$a_i = (x_i - \bar{x})/\sigma_{xx} \qquad i = 1, 2, \ldots, n$$

$$\boxed{T \triangleq \frac{\hat{\beta} - \beta}{\sqrt{s^2 \sum a_i^2}} \sim t_{(n-2)}}$$

régression linéaire

inférence sur $\mathbb{E}(Y)$

Pour toute valeur de x, la valeur attendue (espérence mathématique) de Y associée est $y = \alpha + \beta x$, avec l'estimateur MLE

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

comme $\hat{\alpha} = \bar{y} + \hat{\beta}\bar{x}$

$$\begin{aligned} \hat{\mu} &= \bar{y} + \hat{\beta}(x - \bar{x}) \\ &= \frac{1}{n}\sum y_i + (x - \bar{x})\sum a_i y_i \\ &= \sum \left[\frac{1}{n} + (x - \bar{x})a_i\right] y_i \end{aligned} \tag{2}$$

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \sigma^2 \sum \left[\frac{1}{n} + (x - \bar{x})a_i\right]^2\right)$$

développons

$$\left[\frac{1}{n} + (x - \bar{x})a_i\right]^2 = \frac{1}{n} + \frac{2(x - \bar{x})}{n} \sum a_i + (x - \bar{x})^2 \sum a_i^2$$

comme $\bar{x} = \frac{1}{n} \sum a_i$, on a

$$\sum (x_i - \bar{x}) = \sum x_i - n\bar{x} = 0$$

et donc $\sum a_i = 0$. De plus $\sum a_i^2 = 1/\sigma_{xx}$ on a

$$\sum \left[\frac{1}{n} + (x - \bar{x})a_i\right]^2 = \frac{1}{n} + \frac{(x - \bar{x})^2}{\sigma_{xx}}$$

$$\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2 \sum (1/n - (x - \bar{x})^2/\sigma_{xx})$$

$$\boxed{T' \triangleq \frac{\hat{\mu} - \mu}{\sqrt{s^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sigma_{xx}} \right)}} \sim t_{(n-2)}}$$

régression linéaire

inférence sur α

$$T'' \triangleq \frac{\hat{\alpha} - \alpha}{\sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sigma_{xx}} \right)}} \sim t_{(n-2)}$$

théorème central limite

$X_1, X_2, \ldots,$ séquence de variables i.i.d. de moyenne μ et de variance σ^2

On suppose qu'il existe r avec X_i qui possède une fonction génératrice des moments définie sur $(-r, r)$

$$\bar{X}_n \triangleq \frac{X_1 + X_2 + \ldots + X_n}{n}$$

alors

$$T_n \triangleq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim_{n\to\infty} \mathcal{N}(0,1)$$

théorème central limite

démonstration

$$
\begin{aligned}
M_n(u) &\triangleq \mathbb{E}\left(e^{uT_n}\right) \\
&= \mathbb{E}\left(\exp\left(u\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\right)\right) \\
&= \mathbb{E}\left(\exp\left(u\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}\right)\right)
\end{aligned}
$$

théorème central limite

démonstration

de plus,

$$\frac{\bar{X}_n - \mu}{\sigma} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

soit $Y_i \triangleq \frac{X_i - \mu}{\sigma}$, alors

$$M_n(u) = \mathbb{E}\left(\exp\left(u \, \frac{\sqrt{n}}{n} \sum_{i=1}^{n} Y_i \right) \right)$$

théorème central limite

démonstration

comme les Y_i sont indépendants et de même distribution

$$\begin{aligned} M_n(u) &= \mathbb{E}\left(\exp\left(u\,\frac{\sqrt{n}}{n}\,Y_1\right)\right)\mathbb{E}\left(\exp\left(u\,\frac{\sqrt{n}}{n}\,Y_2\right)\right)\ldots\mathbb{E}\left(\exp\left(u\,\frac{\sqrt{n}}{n}\,Y_n\right)\right) \\ &= M_{Y_1}\left(\frac{u}{\sqrt{n}}\right)M_{Y_1}\left(\frac{u}{\sqrt{n}}\right)\ldots M_{Y_n}\left(\frac{u}{\sqrt{n}}\right) \\ &= \left(M_{Y_1}\left(\frac{u}{\sqrt{n}}\right)\right)^n \end{aligned}$$

en effectuant le développement de Taylor

$$m_Y\left(\frac{u}{\sqrt{n}}\right) = M_Y(0) + \frac{u}{\sqrt{n}}M_Y^{'} + \frac{u^2}{2n}M_Y^{''}(0) + \frac{u^3}{6\,n^{3/2}}\,M_Y^{'''}(u_n) \qquad u_n \in [0; u/\sqrt{n}]$$

théorème central limite

démonstration

$$M_Y\left(\frac{u}{\sqrt{n}}\right) = 1 + \frac{u^2}{2n} + \frac{u^3}{6\,n^{3/2}} M_Y'''(u_n)$$

et donc

$$\begin{aligned} \log M_n(u) &= n \log\left(M_Y\left(\frac{u}{\sqrt{n}}\right)\right) \\ &= n \log\left(1 + \frac{u^2}{2n} + \frac{u^3}{6\,n^{3/2}} M_Y'''(u_n)\right) \end{aligned}$$

Définissons

$$x_n \triangleq \frac{u^2}{2n} + \frac{u^3}{6\, n^{3/2}} M_Y'''(u_n)$$

on a

$$n \log\left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right) = n\log(1+x_n) = n x_n \frac{\log(1+x_n)}{x_n}$$

Comme u_n converge vers 0, $M_Y'''(u_n)$ converge vers $M_Y'''(0)$, et, en conséquence,

$$\lim_{n\to\infty} n\, x_n = \frac{u^2}{2}$$

Comme x_n converge vers 0 lorsque n tend à l'infini

$$\lim_{n\to\infty} \frac{\log(1+x_n)}{x_n} = 1$$

en conséquence

$$\begin{aligned} \lim_{n\to\infty} \log M_n(u) &= \lim_{n\to\infty} n \log\left(M_Y\left(\frac{1}{\sqrt{n}}\right)\right) \\ &= \lim_{n\to\infty} n x_n \frac{\log(1+x_n)}{x_n} \\ &= \frac{u^2}{2} \end{aligned}$$

et en conclusion

$$\lim_{n\to\infty} M_n(u) = e^{u^2/2} \qquad \text{C.Q.F.D.}$$