Solution du devoir maison

Patrick Tardivel, Université de Bourgogne

Exercice 1. *On considère une série statistique double pondérée* $(X, Y) = \{(x_1, y_1, p_1), \dots, (x_n, y_n, p_n)\}$ *avec* $x_i \neq 0$ *pour un certain* $i \in \{1, \dots, n\}$. *On souhaite donner une formule pour la droite des moindres carrés passant par l'origine.*

1. *En minimisant la fonction suivante*

$$\forall a \in \mathbb{R} \quad f(a) = \sum_{i=1}^{n} p_i (y_i - a x_i)^2,$$

donner une formule pour le coefficient directeur a^* *de la droite des moindres carrés passant par l'origine :* $y = a^* x$.

L'indice de masse corporel d'un individu est le quotient du poids, en kilogramme, par la taille, en mètre, au carré. L'indice de masse corporelle d'un individu normal devrait se situer entre $18,5 kg/m^2$ *et* $25 kg/m^2$. *Ainsi, on s'attend à ce que le poids soit approximativement proportionnelle au carré de la taille. Nous allons vérifier cela à l'aide du fichier "taille_poids.Rdata" qui fournit la taille et le poids d'individus ayant 18 ans et vivant à Honk-Kong au moment de l'enquête statistique.*

2. *Ouvrir le fichier "taille_poids.Rdata" (avec la commande Open File sur RStudio) puis représenter le nuage de points du poids* Y *en kilogramme par rapport à la taille au carré* X *en mètre carré.*
3. *Calculer la droite des moindres carrés passant par l'origine du poids* Y *en fonction de la taille au carré* X.
4. *Ajouter sur à la figure précédente la droite des moindres carrés passant par l'origine.*

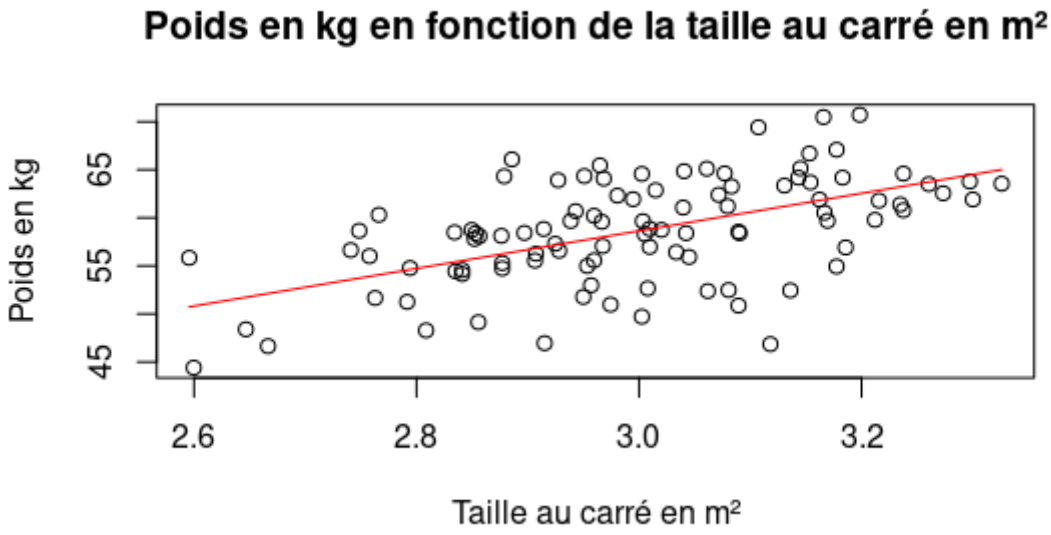
Solution

1. La dérivée de la fonction f vaut

$$\forall a \in \mathbb{R} \quad f'(a) = 2a \sum_{i=1}^{n} p_i x_i^2 - 2 \sum_{i=1}^{n} x_i y_i = 2a\overline{X^2} - 2\overline{XY}$$

Ainsi, $f'(a) = 0$ si et seulement si $a = \overline{XY}/\overline{X^2}$ par ailleurs $f''(a) = 2\overline{X^2} > 0$ donc f est convexe est atteint son minimum au point $a^* = \overline{XY}/\overline{X^2}$.

2, 3 et 4. La droite des moindres carrés passant par l'origine du poids Y en fonction de la taille au carré X a pour équation $y = 19,5469x$. Cette droite ainsi que le nuage de points sont représentés à la figure suivante.



Exercice 2. *Soit $Z_1, \dots, Z_n$ des variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$. On pose $Y = \sum_{i=1}^n Z_i^2$ une variable aléatoire de loi khi-deux à n degrés de liberté. Pour cet exercice, on pourra utiliser les résultats : $\mathbb{E}(Z_1^2) = 1$ et $\mathrm{var}(Z_1^2) = 2$.*

1. *Montrer la convergence en loi suivante*

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{Y-n}{\sqrt{2n}} \leq t\right) = \underbrace{\int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx}_{=\Phi(t)}.$$

   Sachant que $\Phi(1,645) = 0,95$, en déduire une approximation du quantile d'ordre $0,95$ de la loi du khi-deux à $n = 30$ degrés de liberté.

2. *En utilisant la méthode-delta, établir la convergence en loi suivante*

$$\lim_{n\to+\infty} \mathbb{P}\left(\sqrt{2Y} - \sqrt{2n} \leq t\right) = \Phi(t).$$

   En déduire une approximation du quantile d'ordre $0,95$ de la loi du khi-deux à 30 degrés de liberté.

3. *En utilisant le formulaire, déterminer l'approximation la plus précise.*

Solution :

1. Comme l'espérance est linéaire, on a

$$\mathbb{E}(Y/n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Z_i^2) = 1.$$

   Comme les variables aléatoires $Z_1^2, \dots, Z_n^2$ sont indépendantes on a

$$\mathrm{var}(Y/n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(Z_i^2) = 2/n.$$

Ainsi, d'après le théorème de la limite centrale on a la convergence suivante :

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{Y/n-1}{\sqrt{2/n}} \leq t\right) = \lim_{n\to+\infty} \mathbb{P}\left(\frac{Y-n}{\sqrt{2n}} \leq t\right) = \Phi(t).$$

Comme $\Phi(1,645) = 0,95$ on en déduit la limite suivante :

$$\lim_{n\to+\infty} \mathbb{P}\left(Y \leq n + 1,645\sqrt{2n}\right) = 0,95.$$

Donc, pour $n = 30$, une approximation du $0,95$ quantile de la variable aléatoire Y est $30 + 1,645\sqrt{60} = 42,7421$.

2. On pose $g(x) = \sqrt{2x}$ d'où $g'(x) = 1/\sqrt{2x}$. D'après la méthode delta on a

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{g(Y/n)-g(1)}{g'(1)\sqrt{2/n}} \leq t\right) = \lim_{n\to+\infty} \mathbb{P}\left(\sqrt{2Y}-\sqrt{2n} \leq t\right) = \Phi(t).$$

Comme $\Phi(1,645) = 0,95$ on en déduit la limite suivante :

$$\lim_{n\to+\infty} \mathbb{P}\left(\sqrt{2Y} \leq \sqrt{2n}+1,645\right) = \lim_{n\to+\infty} \mathbb{P}\left(Y \leq \frac{(\sqrt{2n}+1,645)^2}{2}\right) = 0,95.$$

Ainsi, pour $n = 30$, une approximation du $0,95$ quantile de la variable aléatoire Y est $0,5(\sqrt{60}+1,645)^2 = 44,0951$.

3. D'après le formulaire, le $0,95$ quantile d'une loi $\chi^2(30 \text{ ddl})$ vaut $43,7729$. La meilleur approximation est celle obtenue avec la méthode delta.