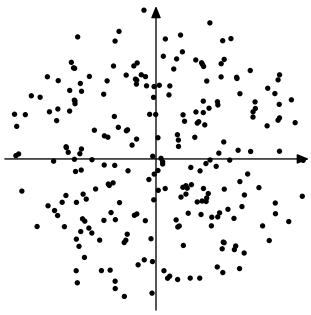
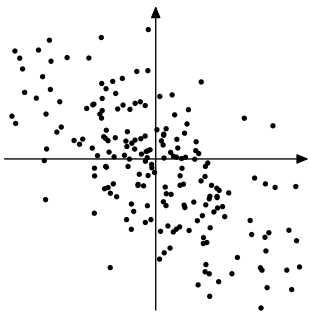
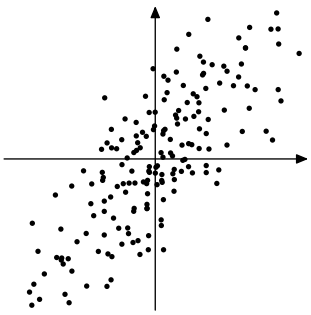
Probabilités et statistique

Série 12

Régression linéaire (solutionnaire)

Exercice 1. Pour chacun des diagrammes de dispersion suivants, indiquer si la corrélation r est strictement positive, strictement négative ou nulle.

a)  b)  c) 

Solution. a) $r = 0$, car il ne semble pas y avoir de relation linéaire entre les deux variables

b) $r < 0$, car la relation linéaire semble avoir une pente négative, c'est-à-dire les points ont tendance à se trouver dans les quadrants II et IV.

c) $r > 0$, car la relation linéaire semble avoir une pente positive, c'est-à-dire les points ont tendance à se trouver dans les quadrants I et III.

Exercice 2. a) Montrer que

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}.$$

b) Montrer que la covariance empirique sans biais vérifie

$$\widehat{\mathrm{cov}}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n} X_i Y_i - \frac{n}{n-1}\overline{X}\,\overline{Y}.$$

Solution. a) Il s'agit d'un calcul direct. On a

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) &= \frac{1}{n}\sum_{i=1}^{n}(X_i Y_i - X_i\overline{Y} - \overline{X}Y_i + \overline{X}\,\overline{Y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{Y}\frac{1}{n}\sum_{i=1}^{n} X_i - \overline{X}\frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{1}{n}\sum_{i=1}^{n}\overline{X}\,\overline{Y} \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{Y}\,\overline{X} - \overline{X}\,\overline{Y} + \overline{X}\,\overline{Y}\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}.$$

b) On peut utiliser le a) pour les simplifier les calculs. On a

$$\begin{aligned}
\widehat{\mathrm{cov}}(X,Y) &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i Y_i - \overline{X}\,\overline{Y}) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i Y_i - \overline{X}\,\overline{Y})\right) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} X_i Y_i - \frac{n}{n-1}\overline{X}\,\overline{Y}
\end{aligned}$$

Exercice 3. Calculer la corrélation empirique de l'échantillon suivant

$$(0,1) \qquad (0.5, 0.25) \qquad (-1,0) \qquad (-0.25, 0.5) \qquad (2,2)$$

Exercice 4. Soit (X,Y) un vecteur aléatoire qui suit une loi normale bidimensionnelle $N(\vec{\mu}, \Sigma)$, où Σ est la matrice de covariance et $\vec{\mu} = (\mu_X, \mu_Y)^T$.

a) On pose $Z = (X - \mu_X, Y - \mu_Y)$. Montrer que Z suit une loi normale bidimensionnelle $N(\vec{0}, \Sigma)$, où $\vec{0}$ est le vecteur nul.

b) Soit ρ la corrélation entre X et Y. Montrer que si $\rho = 0$, alors X et Y sont indépendantes. *Suggestion*. Montrer d'abord que $X - \mu_X$ et $Y - \mu_Y$ sont indépendants et déduire que X et Y le sont.

c) Dans le cas où la corrélation entre X et Y ρ est nulle, déterminer les lois maginales de X et Y.

Solution. a) On rappelle que la fonction de densité de la loi $N(\vec{\mu}, \Sigma)$ est

$$f(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_X^2} - 2\frac{\rho(x-\mu_x)(y-\mu_y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_y)^2}{\sigma_Y^2}\right)\right].$$

Ensuite, pour $Z = (X - \mu_x, Y - \mu_Y)^T$, on a

$$\begin{aligned}
\mathbb{P}(X - \mu_x \le x \text{ et } Y - \mu_y \le y) &= \mathbb{P}(X \le x + \mu_x \text{ et } Y \le y + \mu_y) \\
&= \int_{-\infty}^{x+\mu_x}\int_{-\infty}^{y+\mu_y} f(s,t)\mathrm{d}t\mathrm{d}s \qquad \boxed{\begin{array}{l} u = s - \mu_X \\ v = t - \mu_Y \\ \mathrm{d}u\mathrm{d}v = \mathrm{d}s\mathrm{d}t \end{array}} \\
&= \int_{-\infty}^{x}\int_{-\infty}^{y} f(u + \mu_X, s + \mu_Y)\mathrm{d}v\mathrm{d}u
\end{aligned}$$

et on peut voir que

$$f(u+\mu_X, v+\mu_y) = \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{u^2}{\sigma_X^2} - 2\frac{\rho uv}{\sigma_X\sigma_Y} + \frac{v^2}{\sigma_Y^2}\right)\right]$$

qui est la fonction de densité d'une loi normale $N(\vec{0}, \Sigma)$. Remarquons que $\det\Sigma = \sigma_X^2\sigma_Y^2(1-\rho^2)$, où ρ est la corrélation de X et Y.

b) Supposons que $\rho = 0$. Alors, la fonction de densité de $Z = (X-\mu_X, Y-\mu_Y)^T$ devient

$$\begin{aligned} f_Z(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left(-\frac{x^2+y^2}{2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left(-\frac{y^2}{2}\right) \\ &= f_X(x) f_Y(y). \end{aligned}$$

Cela montre que $X-\mu_X$ et $Y-\mu_Y$ sont indépendantes (et suivent toutes les deux des lois normales). On sait que Y est indépendante de $X-\mu_X$ et que X est indépendante de $Y-\mu_Y$, donc X et Y sont indépendantes.

c) Cela a été fait au b). On voit que Y a la densité de la loi normale $N(\mu_Y,\sigma_Y^2)$ et X, celle de $N(\mu_X,\sigma_X^2)$.

Exercice 5. Calculer la droite de régression de l'échantillon suivant

$$(-1,-1) \qquad (0,1) \qquad (1,0) \qquad (1,-1).$$

Exercice 6. Remarque : il n'y a pas de question de programmation à l'examen, mais peut-être que vous trouverez cette question intéressante, voire amusante.

a) Étant donnée un échantillon observé $(x_1,y_1),\ldots,(x_n,y_n)$, écrire un pseudo-code qui calcule les estimateurs empiriques $\hat{a}$ et $\hat{b}$ dans la régression linéaire $Y = aX + b$.

b) Implémenter le pseudo-code du a) en python ou en R.

Exercice 7. a) On fait le test d'hypothèse $H_0 : a = a_0$, où a est un paramètre de la droite de régression $Y = aX + b$. Pour un échantillon est de taille 18, quelle statistique de test utilise-t-on et quelle est la distribution de celle-ci ? On suppose ici que le résidu est normalement distribuée.

b) Même question que le a), mais pour le test d'hypothèse $H_0 : b = b_0$.

Solution. a) On utilise la statistique de test

$$T_{\text{stat}} = \frac{\hat{a} - a_0}{S_e/\sqrt{(n-2)S_x^2}},$$

où $S_e^2 = \sum_{i=1}^{18} e_i^2$, avec $e_i = Y_i - aX_i - b$, et $S_X^2 = \sum_{i=1}^{18}(X_i - \overline{X})^2$. Sous H_0, T_{stat} suit une loi de Student avec $n-2 = 18-2 = 16$ degrés de liberté.

b) Pour le b, on utilise

$$T_{\text{stat}} = \frac{\hat{b} - b_0}{(\sum X_i^2)\sigma \Big/ \sqrt{n(n-2)S_X^2}}.$$

Sous H_0, T_{stat} suit une loi de Student avec $n - 2 = 18 - 2 = 16$ degrés de liberté.

Exercice 8. On fait l'hypothèse $H_0 : a = 1$. Determiner si l'on peut rejeter H_0 avec un seuil $\alpha = 5\%$, si on observe les données de la question 5. On suppose que le résidu est normalement distribué.

Exercice 9. On fait l'hypothèse $H_0 : b = 2$. Determiner si l'on peut rejeter H_0 avec un seuil $\alpha = 5\%$, si on observe les données de la question 5. On suppose que le résidu est normalement distribué.

Exercice 10. En classe, on a montré que si les résidus $e_1, \ldots, e_n$ sont iid et suivent une loi $N(0,\sigma^2)$, alors les estimateurs $\hat{a}$ et $\hat{b}$ de la régression $Y = aX + b + e$ sont sans biais. Vérifier que $\hat{a}$ et $\hat{b}$ sont encore sans biais si on suppose seulement que $e_1, \ldots, e_n$ sont iid et d'espérance nulle.

Solution. La démarche est la même qu'on a vu en classe. On suppose qu'on a observé $x_1, \ldots, x_n$. Pour $\hat{a}$, on rappelle

$$\hat{a} = \frac{\widehat{\text{cov}}(X,Y)}{\sigma_X^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{s_x^2},$$

où $s_x^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$. On a donc

$$\mathbb{E}(\hat{a}) = \frac{1}{s_x^2}\sum_{i=1}^{n}(x_i - \overline{x})\big(\mathbb{E}[Y_i] - \mathbb{E}[\overline{Y}]\big). \qquad (*)$$

Comme $Y_i = ax_i + b + e_i$, on a $\mathbb{E}(Y_i) = ax_i + b + \mathbb{E}(e_i) = ax_i + b + 0$. De façon similaire, comme $\overline{Y} = a\overline{x} + b + \overline{e}$, on a $\mathbb{E}(\overline{Y}) = a\overline{x} + b + 0$. Il suit que

$$\mathbb{E}(Y_i) - \mathbb{E}(\overline{Y}) = (ax_i + b) - (a\overline{x} + b) = a(x_i - \overline{x}).$$

Si on remplace dans l'équation $(*)$, on obtient

$$\mathbb{E}(\hat{a}) = \frac{1}{s_x^2}\sum_{i=1}^{n}(x_i - \overline{x})a(x_i - \overline{x}) = \frac{a}{s_x^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{a}{s_x^2}s_x^2 = a.$$

Pour $\hat{b}$, c'est plus simple. On rappelle que $\hat{b} = \overline{Y} - a\overline{x}$ et donc on trouve $\mathbb{E}(\hat{b}) = \mathbb{E}(\overline{Y}) - a\overline{x} = (a\overline{x} + b) - a\overline{x} = b$.