Plaquette 5 – Synthèse : études statistiques complètes

Statistiques à deux variables — Première générale

Corrections détaillées

Boîte à outils

Le plan complet d'une étude statistique à deux variables.

— **1. Nuage.** On le construit ou on le lit : allongé ? croissant ou décroissant ? point aberrant ?

— **2. Point moyen.**

$$G(\overline{x}\,;\overline{y}) \qquad \overline{x} = \frac{\sum x_i}{n}\,, \quad \overline{y} = \frac{\sum y_i}{n}$$

— **3. Droite de régression** de y en x : $y = ax + b$, coefficients donnés par la calculatrice, avec le contrôle $\overline{y} = a\overline{x} + b$.

— **4. Corrélation.** r compris entre -1 et 1, du signe de a ; $|r| > 0{,}95$ ajustement très satisfaisant, $|r| < 0{,}8$ ajustement à écarter.

— **5. Estimation.** y connaissant x : on substitue. x connaissant y : on résout, puis on teste les entiers voisins.

— **6. Critique**, en trois questions : l'ajustement affine est-il **légitime** (r, allure du nuage) ? l'estimation est-elle une **interpolation** ou une **extrapolation**, et à quelle distance ? le résultat est-il **possible** (positif, borné, réaliste) ?

— **Interprétation des coefficients.** a s'exprime en unité de y **par** unité de x ; b est la valeur estimée pour $x = 0$, à n'interpréter que si cette situation existe.

— **Corrélation et causalité.** Un $|r|$ proche de 1 ne prouve jamais un lien de cause à effet : il peut exister une variable cachée.

— **La phrase de conclusion.** Elle contient le résultat, son unité, et le degré de confiance. « D'après ce modèle, on estime … ; il s'agit d'une extrapolation de N rangs, à considérer comme une tendance. »

Correction de l'exercice 1 – Étude complète : prix d'un forfait mobile

a) Avec $n = 7$:

$$\overline{x} = \frac{1+2+3+4+5+6+7}{7} = \frac{28}{7} = 4$$

$$\overline{y} = \frac{38+42+47+50+55+59+63}{7} = \frac{354}{7} \approx 50{,}57$$

Le point moyen est environ $G(4\,;50{,}57)$.

b) On remplace x par 4 dans l'équation $y = 4{,}179x + 33{,}857$:

$$y = 4{,}179 \times 4 + 33{,}857 = 16{,}716 + 33{,}857 = 50{,}573$$

On retrouve $\overline{y} \approx 50{,}57$: la droite passe bien par le point moyen ✓

c) Le coefficient $r \approx 0{,}999$ est très proche de 1 : le nuage est presque parfaitement aligné, l'ajustement affine est **très satisfaisant**.

Le coefficient $a \approx 4{,}179$ s'exprime en euros **par an** : le prix moyen du forfait augmente d'environ 4,18 **€ chaque année**.

d) Pour $x = 10$:

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

$$y = 4{,}179 \times 10 + 33{,}857 = 41{,}79 + 33{,}857 \approx 75{,}65$$

Le prix estimé de l'année 10 est donc d'environ 75,65 **€**. Il s'agit d'une **extrapolation** de trois rangs au-delà du dernier relevé (l'année 7), soit environ 50 % de l'étendue des données : elle est plausible, mais suppose que la hausse se poursuive au même rythme.

La réserve de fond. Sur un marché concurrentiel comme la téléphonie, les prix peuvent aussi bien stagner que baisser. Un modèle ajusté sur sept ans ne préjuge pas des décisions commerciales des opérateurs.

Réponse. a) $G(4\,;50{,}57)$ — b) la droite donne 50,57 en $x = 4$ ✓ — c) $r \approx 0{,}999$: excellent ; +4,18 € par an — d) environ 75,65 €, extrapolation de trois rangs.

Correction de l'exercice 2 – Étude complète : pistes cyclables

a) Avec $n = 6$:

$$\overline{x} = \frac{0+1+2+3+4+5}{6} = \frac{15}{6} = 2{,}5 \qquad \overline{y} = \frac{12+18+25+33+38+46}{6} = \frac{172}{6} \approx 28{,}67$$

Le point moyen est $G(2{,}5\,;28{,}67)$. Contrôle :

$$6{,}8 \times 2{,}5 + 11{,}667 = 17 + 11{,}667 = 28{,}667$$

On retrouve $\overline{y}$ ✓

b) Le coefficient $a = 6{,}8$ s'exprime en kilomètres **par an** : la ville aménage en moyenne 6,8 km de pistes cyclables chaque année.

Le coefficient $b \approx 11{,}667$ est la longueur estimée pour $x = 0$, c'est-à-dire en 2019. Ici $x = 0$ correspond à une situation **réelle**, et le relevé de 2019 vaut 12 km : l'estimation 11,7 km est cohérente, l'écart de 0,3 km n'étant qu'un résidu.

c) L'année 2027 correspond à $x = 2027 - 2019 = 8$:

$$y = 6{,}8 \times 8 + 11{,}667 = 54{,}4 + 11{,}667 \approx 66{,}07$$

Soit environ 66 **km** de pistes cyclables. C'est une extrapolation de trois rangs.

d) On résout l'inéquation :

$$6{,}8x + 11{,}667 > 100 \Leftrightarrow 6{,}8x > 88{,}333 \Leftrightarrow x > \frac{88{,}333}{6{,}8} \approx 12{,}99$$

L'objectif serait donc atteint pour $x = 13$, soit en **2032**. Contrôle : $x = 12$ donne 93,3 km (objectif non atteint) et $x = 13$ donne 100,1 km ✓

La réserve. Huit ans d'extrapolation supposent un budget d'aménagement constant, et un réseau qui ne sature pas. Dans les faits, les kilomètres les plus faciles se construisent d'abord : le rythme a plutôt tendance à ralentir.

Réponse. a) $G(2{,}5\,;28{,}67)$ et la droite y passe ✓ — b) 6,8 km par an ; 11,7 km en 2019 — c) environ 66 km en 2027 — d) en 2032 ($x = 13$).

Correction de l'exercice 3 – Étude complète : réduction des déchets

a) Le coefficient r est **négatif**, comme a : la corrélation est négative, la production de déchets **diminue** au fil des années. Sa valeur absolue, 0,9999, est quasiment égale à 1 : le nuage est presque parfaitement aligné et l'ajustement affine est excellent.

b) Le coefficient $a \approx -26{,}036$ s'exprime en kg par habitant **et par an** : la production baisse d'environ 26 **kg par habitant chaque année**.

c) L'année 2028 correspond à $x = 10$:

$$y = -26{,}036 \times 10 + 539{,}25 = -260{,}36 + 539{,}25 = 278{,}89$$

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

Soit environ 279 **kg** par habitant. C'est une extrapolation de quatre rangs, modérée.

d) On résout l'inéquation :

$$-26{,}036x + 539{,}25 < 300 \Leftrightarrow -26{,}036x < -239{,}25 \Leftrightarrow x > \frac{239{,}25}{26{,}036} \approx 9{,}19$$

(la division par un nombre négatif change le sens de l'inégalité). Le seuil serait donc franchi pour $x = 10$, soit en **2028**. Contrôle : $x = 9$ donne 305,0 kg et $x = 10$ donne 278,9 kg ✓

e) Non. Le modèle donne une production **nulle** lorsque

$$-26{,}036x + 539{,}25 = 0 \Leftrightarrow x = \frac{539{,}25}{26{,}036} \approx 20{,}7$$

soit vers 2039, et **négative** ensuite — ce qui est impossible. En réalité, la baisse ralentira : il existe une quantité de déchets irréductible, et les gisements les plus faciles à réduire le sont d'abord. Le modèle affine ne vaut que sur quelques années.

Ce que le correcteur attend en e). Non pas seulement « c'est impossible », mais le **calcul** de l'année où le modèle devient absurde, et le mécanisme qui l'explique.

Réponse. a) r négatif et $|r| \approx 1$: baisse presque parfaitement linéaire — b) environ −26 kg par habitant et par an — c) environ 279 kg en 2028 — d) en 2028 ($x = 10$) — e) non : production nulle vers 2039 ($x \approx 20{,}7$), puis négative.

Correction de l'exercice 4 – Étude complète : arrosage d'un potager

a) Avec $n = 6$:

$$\overline{x} = \frac{10 + 15 + 20 + 25 + 30 + 35}{6} = \frac{135}{6} = 22{,}5$$

$$\overline{y} = \frac{46 + 61 + 78 + 92 + 109 + 123}{6} = \frac{509}{6} \approx 84{,}83$$

Le point moyen est $G(22{,}5\,;\,84{,}83)$. Contrôle :
$3{,}103 \times 22{,}5 + 15{,}019 \approx 69{,}82 + 15{,}019 \approx 84{,}84$ ✓

b) Pour $x = 28$:

$$y = 3{,}103 \times 28 + 15{,}019 \approx 86{,}88 + 15{,}019 \approx 101{,}90$$

soit environ 102 **litres**. C'est une **interpolation** (28 est entre 25 et 30) : l'estimation est fiable.

Pour $x = 50$:

$$y = 3{,}103 \times 50 + 15{,}019 = 155{,}15 + 15{,}019 \approx 170{,}17$$

soit environ 170 **litres**. C'est une **extrapolation** de 15 m² au-delà du dernier relevé, soit 60 % de l'étendue : plausible, mais à annoncer comme telle.

c) Le coefficient $a \approx 3{,}10$ s'exprime en litres **par mètre carré** : chaque mètre carré supplémentaire demande environ 3,1 litres d'eau.

Le coefficient $b \approx 15$ litres est la quantité estimée pour une surface nulle. Elle correspond à une **partie fixe** : l'eau perdue dans le tuyau, l'amorçage, l'arrosage des allées.

La quantité d'eau n'est donc **pas proportionnelle** à la surface : si elle l'était, on aurait $b = 0$. On le vérifie sur les quotients : $\frac{46}{10} = 4{,}6$ L/m² pour le petit potager, contre $\frac{123}{35} \approx 3{,}51$ L/m² pour le grand. Le coût en eau par mètre carré **baisse** avec la surface, précisément à cause de la partie fixe.

Le test de proportionnalité, en une phrase. Deux grandeurs sont proportionnelles si et seulement si le nuage est aligné **et passe par l'origine**. Ici il est aligné, mais son

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

ordonnée à l'origine vaut 15.

Réponse. a) $G(22{,}5\,;\,84{,}83)$ — b) environ 102 L (interpolation) et 170 L (extrapolation) — c) 3,1 L par m² plus une partie fixe de 15 L : non proportionnel.

Correction de l'exercice 5 – Étude complète : salaire et ancienneté

a) Avec $n = 7$:

$$\overline{x} = \frac{0+2+5+8+12+16+20}{7} = \frac{63}{7} = 9$$

$$\overline{y} = \frac{1\,700+1\,790+1\,930+2\,060+2\,250+2\,410+2\,580}{7} = \frac{14\,720}{7} \approx 2\,102{,}86$$

Le point moyen est $G(9\,;\,2\,102{,}86)$. Contrôle :

$$44{,}141 \times 9 + 1\,705{,}587 = 397{,}269 + 1\,705{,}587 \approx 2\,102{,}86$$

On retrouve $\overline{y}$ ✓

b) Le coefficient $a \approx 44{,}14$ s'exprime en euros **par année d'ancienneté** : chaque année d'ancienneté correspond, en moyenne, à environ 44 € de salaire mensuel supplémentaire.

Le coefficient $b \approx 1\,706$ € est le salaire estimé pour une ancienneté **nulle**, c'est-à-dire à l'embauche. Ici $x = 0$ correspond à une situation réelle, et le relevé donne 1 700 € : l'estimation est cohérente à 6 € près.

c) Pour $x = 25$:

$$y = 44{,}141 \times 25 + 1\,705{,}587 \approx 1\,103{,}53 + 1\,705{,}587 \approx 2\,809{,}11$$

Le salaire estimé est donc d'environ 2 809 **€**. C'est une extrapolation de cinq ans au-delà du dernier relevé, soit 25 % de l'étendue : elle reste raisonnable.

d) On résout l'inéquation :

$$44{,}141x + 1\,705{,}587 > 3\,000 \Leftrightarrow 44{,}141x > 1\,294{,}413 \Leftrightarrow x > \frac{1\,294{,}413}{44{,}141} \approx 29{,}32$$

Le seuil serait franchi à partir de 30 **ans** d'ancienneté. Contrôle : $x = 29$ donne 2 985,7 € et $x = 30$ donne 3 029,8 € ✓

La réserve : 30 ans est une extrapolation de dix ans au-delà des données. Surtout, une grille salariale réelle n'est presque jamais affine sur toute une carrière — les augmentations à l'ancienneté sont souvent plafonnées, et les progressions les plus fortes viennent des changements de poste, que ce modèle ignore complètement.

La leçon de modélisation. Un modèle ajusté sur une population à un instant donné décrit une **photographie**, pas une trajectoire individuelle : rien ne dit que le salarié embauché aujourd'hui suivra la même progression.

Réponse. a) $G(9\,;\,2\,102{,}86)$, la droite y passe ✓ — b) environ 44 € par année d'ancienneté ; 1 706 € à l'embauche — c) environ 2 809 € — d) à partir de 30 ans, sous réserve que la grille reste affine.

Correction de l'exercice 6 – Étude complète : boissons chaudes et température

a) Le coefficient $r \approx -0{,}9996$ est négatif et de valeur absolue quasiment égale à 1 : la corrélation est **négative** et l'alignement quasi parfait. L'ajustement affine est excellent sur l'intervalle observé.

Le coefficient $a \approx -1{,}421$ s'exprime en boissons **par degré** : chaque degré supplémentaire fait perdre environ 1,42 **vente** par jour. C'est parfaitement cohérent : plus il fait chaud, moins on achète de boissons chaudes.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

b) Pour $x = 12$:

$$y = -1{,}421 \times 12 + 44{,}929 = -17{,}052 + 44{,}929 \approx 27{,}88$$

Soit environ 28 **boissons** par jour. C'est une **interpolation** (12 est entre 10 et 15) : l'estimation est fiable.

c) Pour $x = 30$:

$$y = -1{,}421 \times 30 + 44{,}929 = -42{,}63 + 44{,}929 \approx 2{,}30$$

soit environ 2 boissons par jour. Le résultat est encore possible, mais c'est une extrapolation de 5 °C au-delà des relevés.

Pour $x = 40$:

$$y = -1{,}421 \times 40 + 44{,}929 = -56{,}84 + 44{,}929 \approx -11{,}91$$

Le modèle prévoit -12 ventes, ce qui est **impossible**. On peut calculer précisément où il cesse d'avoir un sens :

$$-1{,}421x + 44{,}929 = 0 \Leftrightarrow x = \frac{44{,}929}{1{,}421} \approx 31{,}6$$

Au-delà de 31,6 °C, le modèle donne des ventes négatives.

La réalité derrière le modèle. Les ventes ne tombent pas à zéro par forte chaleur : il reste toujours quelques clients. La relation s'aplatit au lieu de traverser l'axe — c'est le comportement typique d'une grandeur **positive** qui tend vers un plancher.

Réponse. a) $r \approx -0{,}9996$: excellent, corrélation négative ; environ $-1{,}42$ vente par degré — b) environ 28 boissons — c) environ 2 boissons à 30 °C, mais -12 à 40 °C : impossible dès 31,6 °C.

Correction de l'exercice 7 – Étude complète : révisions et résultats

a) Avec $n = 8$:

$$\overline{x} = \frac{0 + 1 + 2 + 3 + 4 + 5 + 6 + 7}{8} = \frac{28}{8} = 3{,}5$$

$$\overline{y} = \frac{4 + 6 + 7 + 9 + 12 + 13 + 15 + 16}{8} = \frac{82}{8} = 10{,}25$$

Le point moyen est $G(3{,}5\,;10{,}25)$. Contrôle :

$$1{,}786 \times 3{,}5 + 4 = 6{,}251 + 4 = 10{,}251$$

On retrouve $\overline{y}$ aux arrondis près ✓

b) Le coefficient $a \approx 1{,}786$ s'exprime en points **par heure** de révision : une heure de révision supplémentaire correspond, en moyenne, à environ 1,8 **point** de plus.

Le coefficient $b = 4$ est la note estimée pour $x = 0$, soit un élève qui n'a pas révisé : 4 sur 20. Ici $x = 0$ existe bien dans la série, et le relevé donne exactement 4 : l'ajustement est parfait sur ce point.

c) Pour $x = 4{,}5$:

$$y = 1{,}786 \times 4{,}5 + 4 = 8{,}037 + 4 \approx 12{,}04$$

La note estimée est d'environ 12 **sur** 20. C'est une interpolation, donc une estimation fiable — à la dispersion du nuage près, qui reste faible ici ($r \approx 0{,}994$).

d) On résout :

$$1{,}786x + 4 > 20 \Leftrightarrow 1{,}786x > 16 \Leftrightarrow x > \frac{16}{1{,}786} \approx 8{,}96$$

Dès environ 9 **heures** de révision, le modèle annonce une note supérieure à 20, ce qui

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

est impossible. Son domaine de validité est donc au plus $[0\,;9]$ heures — et en pratique l'intervalle des relevés, $[0\,;7]$.

La raison de fond. Une note est une grandeur **bornée** par 20. Aucun modèle affine croissant ne peut rester valable indéfiniment sur une grandeur bornée : il finit toujours par dépasser la borne. C'est le même phénomène que pour un pourcentage ou un taux de remplissage.

Réponse. a) $G(3{,}5\,;10{,}25)$, la droite y passe ✓ — b) environ 1,8 point par heure ; 4 / 20 sans réviser — c) environ 12 / 20 — d) dès 9 heures, donc validité limitée à $[0\,;7]$ au plus.

Correction de l'exercice 8 – Étude complète : pluie et fréquentation

a) Avec $n = 5$:

$$\overline{x} = \frac{20+40+60+80+100}{5} = \frac{300}{5} = 60 \qquad \overline{y} = \frac{12+9+7+4+2}{5} = \frac{34}{5} = 6{,}8$$

Le point moyen est $G(60\,;6{,}8)$. Contrôle :

$$-0{,}125 \times 60 + 14{,}3 = -7{,}5 + 14{,}3 = 6{,}8$$

On retrouve exactement $\overline{y}$ ✓

b) Le coefficient $a = -0{,}125$ s'exprime en centaines de visiteurs **par millimètre** de pluie : chaque millimètre supplémentaire fait perdre 0,125 centaine de visiteurs, soit 12,5 **visiteurs par millimètre**.

Autrement dit, 10 mm de pluie en plus coûtent environ 125 visiteurs.

c) Pour $x = 70$:

$$y = -0{,}125 \times 70 + 14{,}3 = -8{,}75 + 14{,}3 = 5{,}55$$

Soit environ 555 **visiteurs**. C'est une interpolation (70 est entre 60 et 80) : l'estimation est fiable.

d) Le modèle donne une fréquentation nulle lorsque

$$-0{,}125x + 14{,}3 = 0 \Leftrightarrow x = \frac{14{,}3}{0{,}125} = 114{,}4$$

soit pour 114,4 mm de pluie, et **négative** au-delà. Ce n'est pas réaliste : même sous une pluie diluvienne, quelques visiteurs se présentent (abonnés, groupes réservés), et surtout le parc finirait par fermer — la relation cesse d'être affine bien avant.

L'extrapolation est de plus lointaine : 114 mm dépasse de 14 % la pluviométrie maximale observée, mais la question est surtout que le modèle **franchit zéro**, ce qu'aucune fréquentation ne fait.

Le réflexe à avoir sur toute grandeur positive. Chercher où la droite coupe l'axe des abscisses : cela donne immédiatement la borne du domaine de validité du modèle.

Réponse. a) $G(60\,;6{,}8)$, la droite y passe exactement ✓ — b) environ 12,5 visiteurs de moins par millimètre — c) environ 555 visiteurs — d) nulle pour 114,4 mm, ce qui n'est pas réaliste.

Correction de l'exercice 9 – Comparer deux séries par leur corrélation

a) On compare $|r|$ aux repères du cours.

Étude 1 : $|r| = 0{,}96 > 0{,}95$, l'ajustement affine est **justifié**.

Étude 2 : $|r| = 0{,}62 < 0{,}8$, l'ajustement affine est à **écarter** : le nuage est trop dispersé, la droite n'aurait pas de valeur prédictive.

b) Non, et pour deux raisons distinctes.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

D'abord, r mesure la **qualité de l'alignement**, pas l'ampleur de l'effet : une relation très bien alignée peut avoir un coefficient directeur minuscule, donc un effet négligeable. C'est a, et non r, qui mesure l'effet.

Ensuite, même avec $r \approx 0{,}96$, la corrélation ne prouve pas la **causalité** : il est possible que les périodes de forte demande aient entraîné à la fois un budget publicitaire élevé et des ventes importantes.

c) Il faudrait comparer, pour chaque action, le **gain de ventes par euro dépensé**, c'est-à-dire les coefficients directeurs ramenés à un même coût — et, idéalement, mener une **expérience contrôlée** : faire varier un seul facteur à la fois, sur des zones comparables, et mesurer l'écart de ventes.

La distinction à retenir. r répond à « le modèle affine est-il pertinent ? », a répond à « quel est l'effet, et de quelle ampleur ? ». Confondre les deux est l'erreur d'interprétation la plus fréquente sur ce chapitre.

Réponse. a) l'étude 1 (0,96) ; l'étude 2 (0,62) est à écarter — b) non : r mesure l'alignement, pas l'effet, et une corrélation n'est pas une causalité — c) comparer les effets par euro dépensé, via une expérience contrôlée.

Correction de l'exercice 10 – Retrouver des données à partir de résumés

a) La droite de régression passe toujours par le point moyen, donc

$$\overline{y} = a\overline{x} + b$$

On remplace par les données connues :

$$140 = 4{,}6 \times 25 + b \Leftrightarrow 140 = 115 + b \Leftrightarrow b = 25$$

b) L'équation de la droite est donc

$$y = 4{,}6x + 25$$

Contrôle : pour $x = 25$, $y = 115 + 25 = 140 = \overline{y}$ ✓

Pour $x = 30$:

$$y = 4{,}6 \times 30 + 25 = 138 + 25 = 163$$

L'estimation est 163.

c) Oui. La moyenne est la somme divisée par l'effectif, donc la somme est le produit de la moyenne par l'effectif :

$$\sum x_i = n \times \overline{x} = 10 \times 25 = 250$$

De même, $\sum y_i = 10 \times 140 = 1\,400$.

Ce que l'exercice montre. Les résumés statistiques (n, $\overline{x}$, $\overline{y}$, a, r) contiennent beaucoup d'information : ils suffisent à reconstituer b, les sommes, et donc à faire toutes les estimations. C'est pourquoi les énoncés de bac les fournissent souvent au lieu des données brutes.

La réserve. En revanche, ils ne disent rien de l'**intervalle** des x_i : sans lui, impossible de savoir si l'estimation en $x = 30$ est une interpolation ou une extrapolation. C'est une information qu'il faut réclamer.

Réponse. a) $b = 25$ — b) $y = 4{,}6x + 25$, et $y = 163$ pour $x = 30$ — c) oui, $\sum x_i = 10 \times 25 = 250$.

Correction de l'exercice 11 – Un nuage à deux tendances

a) Le rendement **augmente** d'abord nettement avec la quantité d'engrais, atteint un

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

maximum autour de 140 à 180 kg/ha, puis **diminue** ensuite. Le nuage a la forme d'un arc, pas d'une droite.

b) Non, un ajustement affine sur l'ensemble du nuage n'est **pas pertinent**. Une droite ne peut représenter qu'une seule tendance : ajustée sur ces sept points, elle serait presque horizontale et ne décrirait ni la montée ni la descente. Le coefficient de corrélation serait faible, et surtout la droite passerait **au milieu** du nuage sans en épouser la forme.

On peut le voir autrement : le point moyen est

$$\overline{x} = \frac{20+60+100+140+180+220+240}{7} = \frac{960}{7} \approx 137{,}1 \qquad \overline{y} = \frac{429}{7} \approx 61{,}3$$

et le rendement observé au voisinage de $\overline{x}$ vaut 76, soit bien plus que $\overline{y}$: la droite sous-estimerait le centre du nuage et surestimerait les deux extrémités.

c) Deux solutions, selon l'usage.

Ajuster sur une partie du nuage : sur $[20\,;140]$, la tendance est nettement croissante et un modèle affine y serait légitime — avec un domaine de validité restreint à cette plage.

Changer de modèle : la forme en arc appelle une **parabole** (un trinôme du second degré), qui possède naturellement un maximum. C'est le modèle qu'un agronome utiliserait, car il correspond au phénomène réel : l'engrais améliore le rendement jusqu'à un optimum, au-delà duquel il devient toxique pour la plante.

La compétence évaluée. Savoir **refuser** un ajustement affine. La calculatrice ne refuse jamais : c'est au rédacteur de regarder le nuage avant de calculer une droite.

Réponse. a) le rendement monte, culmine vers 140–180 kg/ha, puis redescend — b) non : une droite ne peut représenter deux tendances opposées — c) ajuster sur $[20\,;140]$, ou changer de modèle pour une parabole.

Correction de l'exercice 12 – Résidus et point mal ajusté

a) On calcule $3x + 10$ pour chaque abscisse, puis le résidu $y_{\text{observé}} - y_{\text{estimé}}$.

x	1	2	3	4	5	6
estimé	13	16	19	22	25	28
observé	13	15	20	22	32	28
résidu	0	−1	+1	0	+7	0

b) Le point le plus mal ajusté est celui d'abscisse 5 : son résidu vaut +7, alors que tous les autres sont compris entre −1 et +1. L'écart est donc **sept fois** supérieur à la dispersion habituelle.

c) La configuration est caractéristique d'un **point aberrant** ou d'une **erreur de relevé**, et non d'un modèle inadapté. Le raisonnement :

— si le modèle était inadapté, les résidus formeraient une **structure** — tous positifs d'un côté, tous négatifs de l'autre, comme sur un nuage en arc ;
— ici, cinq résidus sur six sont quasi nuls : la droite ajuste parfaitement le reste de la série. Seul un point se détache.

La valeur 32 est donc suspecte. On peut supposer une faute de saisie (32 au lieu de 23 ? l'inversion de chiffres est l'erreur la plus courante) : avec 23, le résidu tomberait à −2, dans la norme. La démarche correcte est de **retourner à la source** de la donnée avant de décider de la corriger ou de l'écarter.

Ce que les résidus permettent de diagnostiquer. Des résidus petits et sans structure : bon modèle. Un résidu isolé et grand : point aberrant. Des résidus structurés

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

(tous du même signe par zones) : mauvais modèle.

Réponse. a) résidus 0 ; −1 ; +1 ; 0 ; +7 ; 0 — b) le point d'abscisse 5, avec un résidu de +7 — c) un point aberrant ou une erreur de relevé, car les cinq autres résidus sont quasi nuls.

Correction de l'exercice 13 – Deux variables, deux unités

a) On remplace x par 12 :

$$y = 1{,}35 \times 12 + 4{,}20 = 16{,}20 + 4{,}20 = 20{,}40$$

Le coût estimé est de 20,40 **€**. C'est une interpolation ($12 \in [0{,}5\,;20]$).

b) On résout l'inéquation, en n'oubliant pas que la contrainte est de **ne pas dépasser** 20 € :

$$1{,}35x + 4{,}20 \leq 20 \Leftrightarrow 1{,}35x \leq 15{,}80 \Leftrightarrow x \leq \frac{15{,}80}{1{,}35} \approx 11{,}70$$

On peut donc expédier un colis d'au plus 11,7 **kg**. Contrôle : $1{,}35 \times 11{,}7 + 4{,}2 = 19{,}995$ € ✓ et un colis de 12 kg coûterait 20,40 €, donc trop.

c) Non, le coût n'est pas proportionnel à la masse : l'ordonnée à l'origine 4,20 € n'est pas nulle. Elle représente la **part fixe** — manutention, étiquette, suivi.

Le coût par kilogramme :

$$\text{pour 1 kg: } \frac{1{,}35 + 4{,}20}{1} = 5{,}55 \text{ €/kg}$$

$$\text{pour 20 kg: } \frac{1{,}35 \times 20 + 4{,}20}{20} = \frac{31{,}20}{20} = 1{,}56 \text{ €/kg}$$

Le coût au kilogramme est donc **3,6 fois plus faible** pour le gros colis : c'est l'effet de la part fixe, répartie sur davantage de poids.

L'erreur d'arrondi en b). Arrondir 11,70 à 12 « au plus proche » donnerait un colis à 20,40 €, au-dessus du budget. Dans une contrainte « au plus », on arrondit **vers le bas**.

Réponse. a) 20,40 € — b) au plus 11,7 kg — c) non : 5,55 €/kg pour 1 kg contre 1,56 €/kg pour 20 kg, à cause de la part fixe de 4,20 €.

Correction de l'exercice 14 – Lire un tableau de résultats complet

a) La droite passe par le point moyen, ce qui donne b :

$$\overline{y} = a\overline{x} + b \Leftrightarrow 63 = 6{,}85 \times 8{,}4 + b \Leftrightarrow 63 = 57{,}54 + b \Leftrightarrow b = 5{,}46$$

L'équation de la droite de régression est donc

$$y = 6{,}85x + 5{,}46$$

b) Le coefficient $r = 0{,}91$ est compris entre 0,8 et 0,95 : l'ajustement affine est **acceptable, à commenter**. La tendance est nette, mais les communes s'écartent sensiblement de la droite — ce qui n'étonne pas, le tissu commercial dépendant aussi du tourisme, de la proximité d'une grande ville, du niveau de vie.

Le coefficient $a = 6{,}85$ s'exprime en commerces **par millier d'habitants** : chaque millier d'habitants supplémentaire correspond à environ 7 **commerces** de plus.

c) Pour $x = 12$:

$$y = 6{,}85 \times 12 + 5{,}46 = 82{,}2 + 5{,}46 = 87{,}66 \approx 88 \text{ commerces}$$

Cette estimation est plausible : 12 milliers d'habitants est un peu au-dessus de la moyenne de l'échantillon, donc probablement dans ou près de l'intervalle des données.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

Avec $r = 0{,}91$, l'incertitude reste toutefois notable — l'estimation donne un ordre de grandeur, pas une valeur précise.

Pour $x = 500$:

$$y = 6{,}85 \times 500 + 5{,}46 = 3\,425 + 5{,}46 = 3\,430{,}46$$

L'estimation de 3 430 commerces est une **extrapolation massive** : l'étude porte sur des communes dont la taille moyenne est de 8 400 habitants, et une ville de 500 000 habitants n'a rien de comparable. Le tissu commercial d'une métropole obéit à d'autres logiques (centres commerciaux, spécialisation des quartiers). Cette estimation est **sans valeur**.

L'information manquante. L'énoncé ne donne pas l'intervalle des x_i. Sans lui, on ne peut qualifier rigoureusement la première estimation : c'est une donnée à réclamer avant toute conclusion.

Réponse. a) $y = 6{,}85x + 5{,}46$ — b) $r = 0{,}91$: acceptable à commenter ; environ 7 commerces par millier d'habitants — c) environ 88 commerces (plausible) et 3 430 (extrapolation massive, sans valeur).

Correction de l'exercice 15 – Synthèse finale : location de vélos en libre-service

a) Avec $n = 8$:

$$\overline{x} = \frac{1 + 2 + \ldots + 8}{8} = \frac{36}{8} = 4{,}5$$

$$\overline{y} = \frac{2{,}1 + 2{,}6 + 3{,}2 + 3{,}6 + 4{,}3 + 4{,}8 + 5{,}2 + 5{,}9}{8} = \frac{31{,}7}{8} = 3{,}9625$$

Le point moyen est $G(4{,}5\,;3{,}96)$.

b) Pour les quatre premiers mois :

$$\overline{x}_1 = \frac{1 + 2 + 3 + 4}{4} = 2{,}5 \qquad \overline{y}_1 = \frac{2{,}1 + 2{,}6 + 3{,}2 + 3{,}6}{4} = \frac{11{,}5}{4} = 2{,}875$$

Pour les quatre derniers :

$$\overline{x}_2 = \frac{5 + 6 + 7 + 8}{4} = 6{,}5 \qquad \overline{y}_2 = \frac{4{,}3 + 4{,}8 + 5{,}2 + 5{,}9}{4} = \frac{20{,}2}{4} = 5{,}05$$

Le coefficient directeur de (G_1G_2) :

$$m = \frac{5{,}05 - 2{,}875}{6{,}5 - 2{,}5} = \frac{2{,}175}{4} = 0{,}543\,75$$

L'équation : $y = 0{,}543\,75(x - 2{,}5) + 2{,}875$, soit

$$y = 0{,}543\,75x - 1{,}359\,375 + 2{,}875 = 0{,}544x + 1{,}516$$

c) Pour la droite de régression, on remplace x par 4,5 :

$$y = 0{,}537 \times 4{,}5 + 1{,}546 = 2{,}4165 + 1{,}546 = 3{,}9625$$

On retrouve exactement $\overline{y} = 3{,}9625$ ✓

Les deux droites sont très proches : $y = 0{,}537x + 1{,}546$ contre $y = 0{,}544x + 1{,}516$. Pour $x = 8$, elles donnent 5,84 et 5,87 milliers de locations, soit un écart d'environ 30 locations sur près de 6 000. Avec $r \approx 0{,}999$, le nuage est si bien aligné que les deux méthodes se rejoignent.

d) Pour $x = 10$:

$$y = 0{,}537 \times 10 + 1{,}546 = 5{,}37 + 1{,}546 \approx 6{,}92$$

soit environ 6 920 locations. **Extrapolation** de deux rangs seulement (soit 25 % de l'étendue) : elle est plausible.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

Pour $x = 30$:

$$y = 0{,}537 \times 30 + 1{,}546 = 16{,}11 + 1{,}546 \approx 17{,}66$$

soit environ 17 660 locations. **Extrapolation lointaine** de 22 rangs, près de trois fois l'étendue des données : le résultat ne vaut que comme ordre de grandeur, et probablement pas même cela.

e) Le nombre maximal de locations quotidiennes est

$$6\,000 \times 5 = 30\,000 \text{ locations, soit } 30 \text{ milliers}$$

On résout donc :

$$0{,}537x + 1{,}546 \leq 30 \Leftrightarrow 0{,}537x \leq 28{,}454 \Leftrightarrow x \leq \frac{28{,}454}{0{,}537} \approx 52{,}99$$

Le modèle reste physiquement possible jusqu'au **mois** 52, et devient impossible dès le mois 53. Contrôle : $x = 52$ donne 29,47 milliers (possible) et $x = 53$ donne 30,01 milliers (au-delà de la capacité) ✓

Mais cette borne physique est très loin derrière la borne **statistique** : le modèle cesse d'être crédible bien avant, dès le douzième ou treizième mois, faute de données pour le soutenir. En pratique, la croissance ralentira dès que la population d'usagers potentiels sera équipée.

Le bilan du chapitre. Une étude statistique à deux variables se conduit en six étapes : nuage, point moyen, droite, corrélation, estimation, critique. Les cinq premières se calculent ; la sixième s'écrit — et c'est elle qui distingue une réponse d'un simple résultat.

Réponse. a) $G(4{,}5\,;3{,}96)$ — b) $G_1(2{,}5\,;2{,}875)$, $G_2(6{,}5\,;5{,}05)$ et $y = 0{,}544x + 1{,}516$ — c) la régression passe par G ; les deux droites sont quasi confondues — d) environ 6 920 locations au mois 10 (extrapolation courte) et 17 660 au mois 30 (extrapolation lointaine) — e) jusqu'au mois 52 physiquement, mais la crédibilité statistique s'arrête bien avant.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).