Méthodes statistiques
Solutions Chapitre 12

Quelques commandes Excel permettant d'effectuer les calculs sont présentées à la fin de ce document

Problème 12.1

a) $b_1 = \frac{s_{xy}}{s_x^2} = 0{,}1377551$, $b_o = \bar{y} - b_1\bar{x} = 29{,}69898$. La droite de régression est donc $y =$

b) $r = \frac{s_{xy}}{s_x s_y} = 0{,}2410714$; $\hat{\sigma}_{y.x} = \sqrt{\frac{n-1}{n-2}}\, s_y \sqrt{1-r^2} = 7{,}892399$; $\hat{\sigma}_{b_1} = \frac{\hat{\sigma}_{y.x}}{\sqrt{\sum (x_i - \bar{x})^2}} = 0{,}1012512$

c) La valeur de Z est $Z = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}} = 1{,}360528$. Cette valeur (interprétée comme une cote Z) n'étant pas excessive, il n'est pas évident que la pente β_1 de la droite n'est pas nulle. On ne peut pas affirmer avec confiance qu'il y a une relation entre le temps passé dans l'établissement et le montant dépensé.

d) On estime le montant moyen μ des dépenses pour 60 minutes passées dans l'établissement : $\hat{\mu} = b_o + b_1(60) = 37{,}96$ \$; l'écart-type de l'estimateur est estimé par $\hat{\sigma}_{\hat{\mu}} = \hat{\sigma}_{y.x}\sqrt{\frac{1}{n} + \frac{(60-\bar{x})^2}{\sum (x_i - \bar{x})^2}} = 1{,}5649$. Un intervalle de confiance pour cette moyenne est donné par $\hat{\mu} - 2\,\hat{\sigma}_{\hat{\mu}} \le \mu \le \hat{\mu} + 2\,\hat{\sigma}_{\hat{\mu}}$, soit $34{,}83\ \$ \le \mu \le 41{,}09\ \$$.

[*Remarque* : Nous avons estimé cette moyenne en supposant, contrairement à la conclusion en c), qu'il existe une relation entre le temps passé à l'établissement et le montant dépensé. Il n'y a pas de mal à cela. Il faut se rappeler qu'en c) nous n'avons pas conclu avec confiance que β_1 est nul : nous avons seulement été incapables de conclure que $\beta_1 \neq 0$. Si on ne se servait pas de la variable x, on aurait estimé μ par la moyenne des y, soit 37 \$. Dans la mesure où la relation est faible, l'utilisation de la variable x ne changera pas grand-chose : elle nous a fait estimer μ par 37,96 \$].

Questions supplémentaires

1. Déterminer un intervalle de confiance pour β_1

L'intervalle de confiance pour β_1 est donné par $b_1 - 2\hat{\sigma}_{b_1} \le \beta_1 \le b_1 - 2\hat{\sigma}_{b_1} \Rightarrow -0{,}065 \le \beta_1 \le 0{,}340$. Remarquez que l'intervalle contient la valeur 0, ce qui veut dire que nous ne pouvons pas rejeter l'hypothèse que $\beta_1 = 0$. Cette procédure (consistant à déterminer un intervalle de confiance pour β_1) et la conclusion est exactement équivalente à celle déterminée en c).

2. Calculer $\frac{b_1}{\hat{\sigma}_{b_1}}$ et comparer avec Z.

$\frac{b_1}{\hat{\sigma}_{b_1}} = \frac{0{,}1377551}{0{,}1012512} = 1{,}360528 = Z$. Ce n'est pas une coïncidence : la formule de Z est équivalente à $\frac{b_1}{\hat{\sigma}_{b_1}}$, qui est en fait une cote Z, calculée sous l'hypothèse que $\beta_1 = 0$. Rappelons que la cote Z de b_1 est $\frac{b_1 - E(b_1)}{\sigma_{b_1}} = \frac{b_1 - \beta_1}{\sigma_{b_1}}$. Si on suppose que $\beta_1 = 0$ et on estime σ_{b_1} par $\hat{\sigma}_{b_1}$, on obtient $\frac{b_1}{\hat{\sigma}_{b_1}}$.

Problème 12.2

a) $r = 0{,}01$; $Z = 3$ C_1 : La relation entre le nombre de pièces produites et le taux de défectuosités est extrêmement faible, ce qui signifie que les prédictions seront sujettes à de grossières erreurs. Le fait que Z est grand signifie qu'on peut néanmoins affirmer avec confiance que la relation observée dans l'échantillon est également présente dans la population. La raison est que l'échantillon est assez grand pour que même une faible relation dans l'échantillon ne peut pas être attribuée au hasard.

b) $r = 0{,}95$; $Z = 3$ C_3 sous toutes réserves. : La relation entre le nombre de pièces produites et le taux de défectuosités est élevé, et elle n'est pas accidentelle. On peut affirmer avec confiance que la relation observée dans l'échantillon est également présente dans la population. On pourra faire de bonnes prédictions si le coefficient de corrélation réel est vraiment de l'ordre de grandeur observé. On ne peut pas en être sûr, cependant, car la combinaison de valeurs ici ne peut être obtenue que dans un petit échantillon. Par conséquent le coefficient de corrélation de la population pourrait facilement être inférieur à celui observé.

c) $r = 0{,}95$; $Z = 1$ C_5 La dépendance observée dans l'échantillon est probablement accidentelle. On ne peut pas affirmer que le taux de défectuosités dépend du nombre de pièces fabriquées.

d) $r = -0{,}58$; $Z = 3$ C_4 Le signe de r est le même que celui de Z. Il est impossible que r soit négatif et Z positif.

e) $r = -0{,}95$; $Z = -3$ Même réponse qu'en *b*)

f) $r = 0{,}95$; $Z = 0{,}9$ C_4 Une telle combinaison de valeurs est impossible lorsque n est entier et supérieur à 2.

Questions supplémentaires

g) $r = 0{,}2$, $Z = 3{,}2$ C_1 La relation n'est pas forte dans l'échantillon, mais elle est significative dans le sens qu'on peut dire qu'elle existe réellement dans la population.

h) $r = 0{,}5$, $Z = 1{,}7$ C_2 La relation est meilleure que ci-dessus, mais elle n'est pas significative : elle pourrait bien être due au hasard.

Problème 12.3

a) $r = \dfrac{s_{xy}}{s_x s_y} = 0{,}3095304$

b) $b_1 = \dfrac{s_{xy}}{s_x^2} = 91{,}82736$, $b_o = \bar{y} - b_1\bar{x} = 3936{,}842$.

c) Le restaurateur affirme, en fait, que $\beta_1 = 0$. C'est ce que nous allons tester. On calcule $Z = \dfrac{\sqrt{n-2}\, r}{\sqrt{1-r^2}} = 1{,}72$. On ne peut donc pas rejeter l'affirmation du restaurateur.

d) Si $n = 45$ et $r = 0{,}3095304$, alors $Z = 2{,}135$, et on peut conclure avec un certain degré de confiance que $\beta_1 > 0$—donc que les recettes sont liées au nombre de bouteilles servies.

Problème 12.4

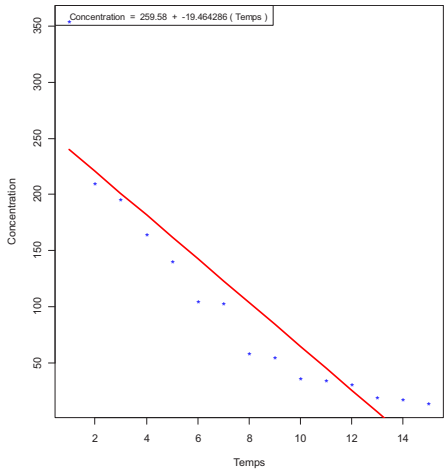
- La valeur de b_1 telle quelle ne suffit pas, car elle dépend des unités de mesures. b_1 est l'accroissement de y correspondant à un accroissement d'une unité de x. Si les unités de x sont petites, l'accroissement de y sera faible.
- C'est exact. L'échantillon pourrait cependant présenter une relation non linéaire.
- Faux. Le coefficient de corrélation r mesure la dépendance présentée par l'échantillon. Dans un petit échantillon une dépendance peut se présenter par hasard.
- Non. Si r est négatif mais assez élevé et l'échantillon assez grand de sorte que $|Z|$ soit grand, alors on conclut que la teneur en fer affecte la corrosion mais de façon contraire à ce qu'on pourrait croire : plus il y a du fer, moins forte est la corrosion.
- Non. Si la valeur absolue de Z est élevée, on peut conclure que la corrélation est significative, c'est-à-dire qu'elle reflète une réelle dépendance, mais ne permet pas de conclure que la dépendance est forte.
- C'est exact.
- C'est exact. Le fait que $|Z|$ est petit signifie que la dépendance *dans la population* peut bien être nulle. Dans le langage habituel, on dit que la corrélation observée est *non significative*.
- C'est exact.

Problème 12.5

- Le coefficient de corrélation, malgré sa faiblesse, permet de conclure qu'il existe une dépendance dans la population. Ceci grâce à un échantillon suffisamment grand. Donc il est vrai que dans la population, ceux qui sont forts en mathématiques ont tendance à être forts en maths. Mais avec $r = 0{,}1$, il est certain que parmi les faibles en maths on trouvera un bon nombre de forts en français (de même que beaucoup de faibles en français parmi les forts en maths).
- C'est exact.
- Non. Bien qu'elle soit faible, elle n'est pas accidentelle. On peut affirmer que si on se limite aux forts en maths, on a de meilleures—très légèrement meilleures—chances de trouver des forts en français (et de moins bonnes chances d'en trouver parmi les faibles en maths.)
- C'est exact. Le hasard tout seul ne peut expliquer la dépendance observée dans l'échantillon, compte tenu de la taille de l'échantillon.
- Non, loin de là. Mais on peut mieux prédire—mais *à peine* mieux— la valeur de y si on connaît la valeur de x que si on ne la connaît pas.
- Non. La valeur de r ne représente pas un niveau de confiance.

Problème 12.6

a)



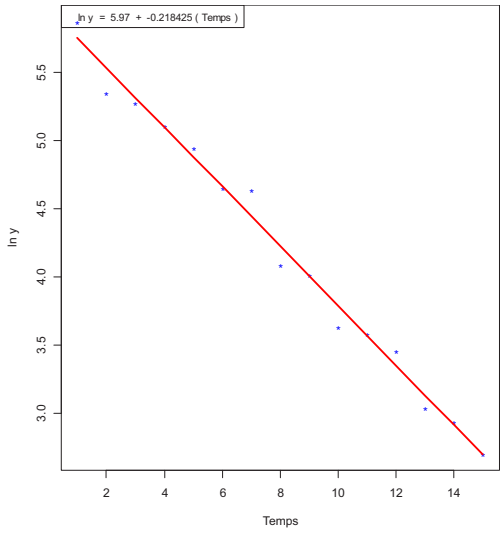
On voit bien, dans le graphique, que la relation n'est pas linéaire : un grand nombre de points successifs au-dessous de la droite de régression, suivis par quatre points, tous au-dessus de la droite.
Soit x le temps et y la concentration.
Quelques calculs : $n = 15$; $\bar{x} = 8$; $\bar{y} = 103{,}8667$; $s_x = 4{,}472136$; $s_y = 95{,}9277$; $s_{xy} = -389{,}2857$; $r = --0{,}9074223$;

$b_1 = \frac{s_{xy}}{s_x^2} = -19{,}46429$; $b_o = \bar{y} - b_1\bar{x} = 259{,}581$; $\hat{\sigma}_{y.x} = \sqrt{\frac{n-1}{n-2}}\, s_y \sqrt{1-r^2} = 41{,}83243$; $Z = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}} = -7{,}785821$; $\hat{\sigma}_{b_1} = 2{,}499966$.
Droite de régression : $y = 259{,}58 - 19{,}46x$.

Le coefficient de corrélation, $r = \frac{s_{xy}}{s_x s_y} = -0{,}9074223$; $\hat{\sigma}_{y.x} = 41{,}83243$; $\hat{\sigma}_{b_1} = 2{,}499966$; $Z = -7{,}785821$.

b) Supposons que le phénomène soit assez bien connu pour savoir que la relation entre N et t est de la forme $N = \beta_o e^{\beta_1 |t|!}$, d'où $\ln N = \ln \beta_o + \beta_1 |t| + \varepsilon$ peut être considéré comme modèle pour $\ln N$. L'ajustement a-t-il l'air meilleur ?

b) Voici le nuage de points :



L'ajustement semble meilleur, en effet.
Quelques calculs
Soit $\delta_o = \ln \beta_o$, d_o l'estimateur de δ_o et $w = \ln y$. $\bar{w} = 4{,}225758$; $s_w = 0{,}9825634$; $s_{xw} = -4{,}368505$; $r = -0{,}9941623$; $b_1 = -0{,}2184253$; $d_o = 5{,}97316$.
Le coefficient de corrélation est ici plus élevé, reflétant un ajustement légèrement meilleur. On peut confirmer que la relation n'est pas accidentelle : $\hat{\sigma}_{w.x} = 0{,}110016$; $\hat{\sigma}_{b_1} = 0{,}006574714$; $Z = -33{,}22202$. Il est certain que la relation observée dans l'échantillon n'est pas accidentelle.
La droite de régression est $w = 5{,}97316 - 0{,}2184253x$.

c) Puisque $\delta_o = \ln \beta_o$ est estimé par $d_o = 5{,}96316$, $\beta_o = e^{\delta_o}$ peut être estimé par $e^{d_o} = 392{,}7449$

d) On estime la moyenne $\mu_{w.8}$ des logarithmes des concentrations par $\hat{\mu} = \hat{\mu}_{w.8} = b_o + b_1(8) = 4{,}225758$ et donc la concentration par $e^{4{,}225758} = 68{,}42637$.
Pour déterminer un intervalle de confiance pour la concentration moyenne, on commence par déterminer un intervalle de confiance pour la moyenne des logarithmes. On a : $\hat{\sigma}_{w.x} = 0{,}110016$; $\hat{\sigma}_{b_1} = 0{,}006574714$ et $\hat{\sigma}_{\hat{\mu}} = 0{,}028406$. L'intervalle de confiance pour la valeur moyenne du logarithme de la concentration est $4{,}168946 \le \mu_{w.8} \le 4{,}282570$. Si $\mu_{w.8}$ se situe entre ces deux bornes, alors la concentration $e^{\mu_{w.8}}$ doit se situer entre $e^{4{,}168946} \le e^{\mu_{w.8}} \le e^{4{,}282570} \Rightarrow 64{,}65 \le e^{\mu_{w.8}} \le 72{,}43$.
[*Remarque* Dans les développements ci-dessus, nous avons pris un certain raccourci que nous nous devons de souligner, bien que la validité des résultats n'en est pas sérieusement compromise. Rappelons que $\mu_{w.8}$ est la moyenne des w, les logarithmes, alors que nous cherchons un intervalle de confiance pour $\mu_{y.8} = E(y \mid x = 8) = E(e^w \mid x = 8)$. Nous avons déterminé un intervalle de confiance pour $E(e^w)$, ce qui n'est pas égal à $e^{E(w)}$.]

Problème 12.7

Quelques calculs : $\bar{x} = 10{,}73333$; $\bar{y} = 5090$; $s_x = 3{,}583327$; $s_y = 983{,}4317$; $s_{xy} = 3317{,}931$.

a) $b_1 = \frac{s_{xy}}{s_x^2} = 258{,}4012$; $b_o = \bar{y} - b_1\bar{x} = 2316{,}494$;

b) $\hat{\sigma}_{y.x} = \sqrt{\frac{n-1}{n-2}}\, s_y \sqrt{1-r^2} = 337{,}1962$; $\hat{\sigma}_{b_1} = \frac{\hat{\sigma}_{y.x}}{\sqrt{\sum (x_i - \bar{x})^2}} = 17{,}4742$

c) $r = 0{,}9415$; $Z = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}} = 15{,}05$. On peut affirmer avec confiance qu'il existe bel et bien une relation entre le poids du courrier et le nombre de commandes.

d) $\hat{\mu}_{y.11,2} = b_o + b_1(11,2) = 5210{,}587$

e) $\hat{\sigma}_{\hat{\mu}} = \hat{\sigma}_{y.x}\sqrt{\frac{1}{n} + \frac{(60-\bar{x})^2}{\sum (x_i - \bar{x})^2}} = 62{,}10104$

Intervalle de confiance pour le nombre moyen de commandes lorsque le poids du courrier est de 11,2 kilos :
$\hat{\mu}_{y.11,2} - 2\,\hat{\sigma}_{\hat{\mu}} \le \mu_{y.11,2} \le \hat{\mu}_{y.11,2} + 2\,\hat{\sigma}_{\hat{\mu}} \Rightarrow 5086 \le \mu_{y.11,2} \le 5335$.

Problème 12.8

a) Les trois coefficients de corrélation sont 0,9671194; 0,5982497; 0,9767395

b) $b_1 = 0{,}05449038$; $b_o = -25{,}2417$

c) $\hat{\sigma}_{y.x} = 3{,}585014$; $\hat{\sigma}_{b_1} = 2{,}348969$

d) $Z = 24{,}94911$. On peut affirmer avec confiance qu'il existe réellement une relation entre le volume de bois et le produit du diamètre par la hauteur.

e) $\hat{\mu} = \hat{\mu}_{y.1000} = 29{,}24867$

f) $\hat{\sigma}_{\hat{\mu}} = 0{,}6449843$

Questions supplémentaires

Déterminer le coefficient de variation des estimateurs $\hat{\mu}$, $\hat{\mu}_1$ et $\hat{\mu}_2$ où :

$\hat{\mu}$ est l'estimateur de $\mu_{y.1000}$ à partir de x (le produit du diamètre par la hauteur est égal) dont la valeur de x est 1000

$\hat{\mu}_1$ est l'estimateur de $\mu_{y.1000}$ à partir de x_1 (le diamètre) dont le diamètre est de 13 pouces

$\hat{\mu}_2$ est l'estimateur de $\mu_{y.1000}$ à partir de x_2 (la hauteur) dont la hauteur est de 76 pieds

(Les valeurs 1000, 13 et 76 sont à peu près les moyennes des variables x, x_1 et x_2.)

$\hat{\mu} = 29{,}24867$; $\hat{\mu}_1 = 28{,}91267$; $\hat{\mu}_2 = 30{,}17097$

$\hat{\sigma}_{\hat{\mu}} / \hat{\mu} = 0{,}02205174$; $\hat{\sigma}_{\hat{\mu}_1} / \hat{\mu}_1 = 0{,}02649867$; $\hat{\sigma}_{\hat{\mu}_2} / \hat{\mu}_2 = 0{,}07975114$

Déterminer la largeur de l'intervalle de confiance de $\mu_{y.1000}$ basé sur chacun des trois estimateurs.

	Limite inférieure	Limite supérieure	Largeur
$\hat{\mu}$	27,96	30,54	2,58
$\hat{\mu}_1$	27,38	30,44	3,06
$\hat{\mu}_2$	25,36	34,98	9,62

Ceci vient confirmer que x est la variable qui prédit le mieux le volume, bien que son avantage par rapport à x_1 est négligeable. L'un ou l'autre est nettement préférable à x_2.

Problème 12.9

a) $b_1 = -443{,}84$; $b_o = 3461{,}052$

b) $r = = -0{,}5391621$

c) $Z = -4{,}48127$

d) $\hat{\mu} = 1463{,}772$; $\hat{\sigma}_{y.x} = 194{,}9141$; $\hat{\sigma}_{\hat{\mu}} = 82{,}98636$; Intervalle de confiance : $1297{,}799 \le \mu_{y.4,5} \le 1629{,}744$

Erratum : Remplacer $y = \gamma_o + \gamma_1 x_1$ par $y = \gamma_o + \gamma_1 x_2$

e) $b_1 = 79{,}51867$; $b_o = 837{,}1714$

f) $r = 0{,}569782$

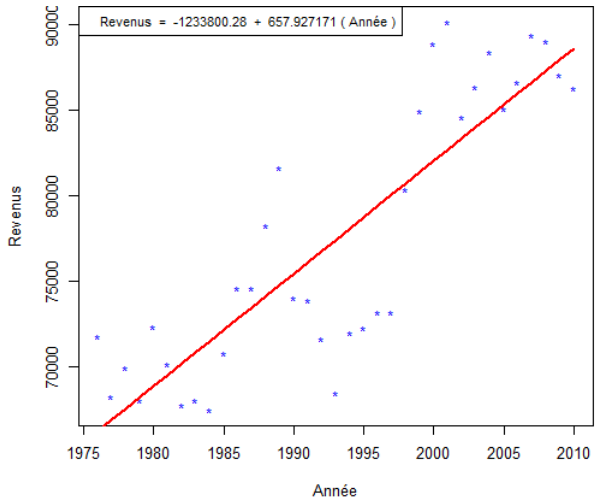
g) $Z = 4{,}853363$

h) $\hat{\sigma}_{y.x} = 190{,}1915$; $\hat{\mu} = 1234{,}765$; $\hat{\sigma}_{\hat{\mu}} = 36{,}4497$

Intervalle de confiance : $1161{,}865 \le \mu_{y.5} \le 1307{,}664$

Problème 12.10

a) Voici le nuage de points :

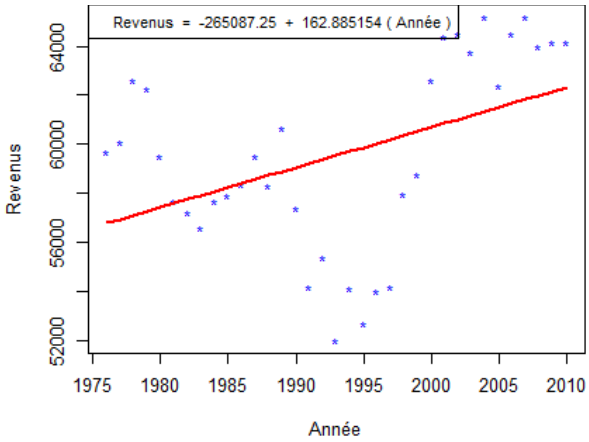


$b_1 =$ 657,9272; $b_o =$ -1 233 800. La relation est donc $y =$ -1 233 800 + 657,9272(Année).
[Nous avons conservé les valeurs de la variable « Année » telles quelles. Il aurait été permis de les remplacer par une séquence de chiffres moins encombrante, comme, par exemple, 0 à 34.]

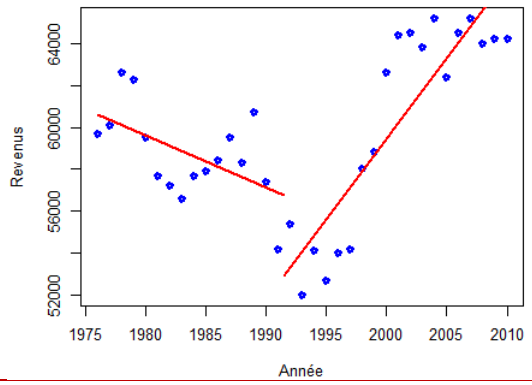
Questions supplémentaires

Peut-on conclure qu'il y a vraiment une tendance (croissante ou décroissante)?
$r =$ 0,8519041; $Z =$ 9,485263. On peut conclure avec confiance qu'il y a une tendance croissante
Déterminer un intervalle de confiance pour la valeur attendue en 2011
Il s'agit de déterminer un intervalle de confiance pour $\mu_{y.2011}$
$\hat{\mu} =$ 89291,26; $\hat{\sigma}_{y.x} =$ 4206,734; $\hat{\sigma}_{\hat{\mu}} =$ 1453,168; Intervalle de confiance : $86\,384{,}93 \leq \mu_{y.2011} \leq 92\,197{,}60$

b) Il est clair que la tendance n'est pas constante sur la durée entière : elle paraît décroissante jusqu'à 1991, puis croissante ensuite.



Montréal de 1976 à 1991 : $b_1 =$ -246,7647; $b_o =$ 548195,3; $r =$ = -0,5522803; $Z =$ -2,565766. La tendance est visiblement négative et statistiquement significative.
Montréal de 1992 à 2010 : $b_1 =$ 771,7544; $b_o =$ -148 405 9; $r =$ 0,8788036; $Z =$ 7,813446. La tendance est visiblement positive, assez forte et statistiquement significative :



Questions supplémentaires

Quelle aurait été, en 1991, l'estimation de $\mu_{y.1992}$ (les revenus attendus de 1992) ? Déterminer un intervalle de confiance pour $\mu_{y.1992}$.

$\hat{\sigma}_{y.x} = 1835{,}638$; $\hat{\sigma}_{b_1} = 99{,}55145$; $\hat{\mu} = 56640$; $\hat{\sigma}_{\hat{\mu}} = 962{,}6167$.

Intervalle de confiance: $54\,714{,}77 \leq \mu_{y.1992} \leq 58\,565{,}23$

Estimer $\mu_{y.2011}$ (les revenus attendus de 2011); déterminer un intervalle de confiance pour $\mu_{y.2011}$.

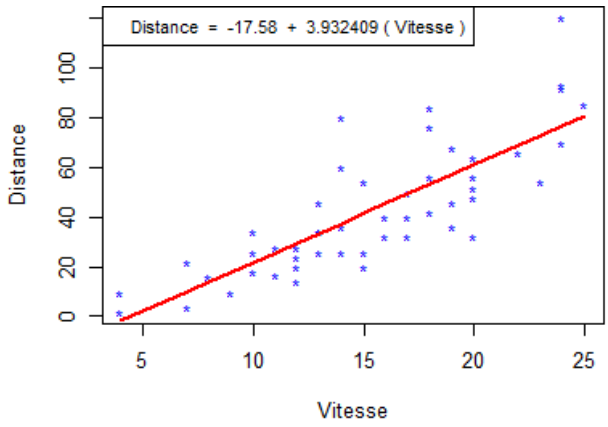
$\hat{\sigma}_{y.x} = 2426{,}53$; $\hat{\sigma}_{b_1} = 20\,3374{,}7$; $\hat{\mu} = 67\,938{,}6$; $\hat{\sigma}_{\hat{\mu}} = 1158{,}831$;

Intervalle de confiance : $65\,620{,}94 \leq \mu_{y.2011} \leq 70\,256{,}26$.

[*Remarque* Si l'analyse avait été basée sur les données de la période entière, l'estimation aurait été $\hat{\mu} = 62\,475$ et l'intervalle de confiance aurait été $59\,897 \leq \mu_{y.2011} \leq 64\,974$. Nous avons rejeté ce modèle (basé sur la période 1976-1991) en faveur d'un modèle linéaire basé sur les années 1992-2010. D'autres choix auraient été possiblement justifiables. Le fait est que la modélisation, qui comprend le choix de la période sur laquelle baser nos prédictions, est un procédé essentiellement subjectif dans lequel l'intuition et l'expérience jouent un rôle important. Le modèle qui en découle est utilisé tant qu'il correspond à la réalité observée; ce n'est pas un fait, c'est une hypothèse qui doit être régulièrement remis en question et ajusté en fonction des observations.]

Problème 12.11

a)



b) $b_1 = 3{,}932409$; $b_o = -17{,}57909$. Droite de régression : Distance = -17,58 + 3,93(Vitesse)

c) $r = 0{,}8068949$. Une relation assez forte laissant prévoir de bonnes prédictions de la distance à partir de la vitesse.

d) $\hat{\sigma}_{y.x} = 15{,}37959.y$

e) $\hat{\sigma}_{b_1} = 0{,}4155128$.

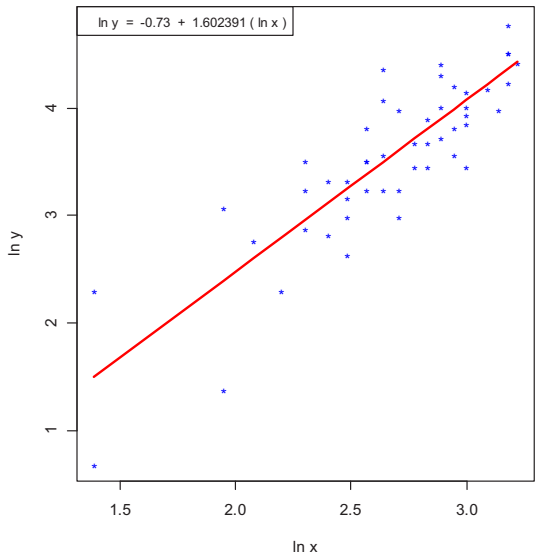
f) $b_1 - 2\hat{\sigma}_{b_1} \leq \beta_1 \leq b_1 + 2\hat{\sigma}_{b_1} \Leftrightarrow 3{,}10 \leq \beta_1 \leq 4{,}76$

g) $r = 0{,}8069$; $Z = 9{,}56$

h) $\hat{\mu} = \hat{\mu}_{y.15} = 41{,}40704$. Intervalle de confiance : $37{,}04435 \leq \mu_{y.15} \leq 45{,}76972$

Problème 12.12

a)



b) $b_1 = = 1{,}602475$; $b_o = -0{,}7298169$. Droite de régression : ln (Distance) = -0,730 + 1,602(ln(Vitesse))

c) $r = 0{,}8562459$. Le coefficient de corrélation est légèrement supérieur ici à celui obtenu au numéro précédent.

d) $\hat{\sigma}_{y.x} = 0{,}4052664$

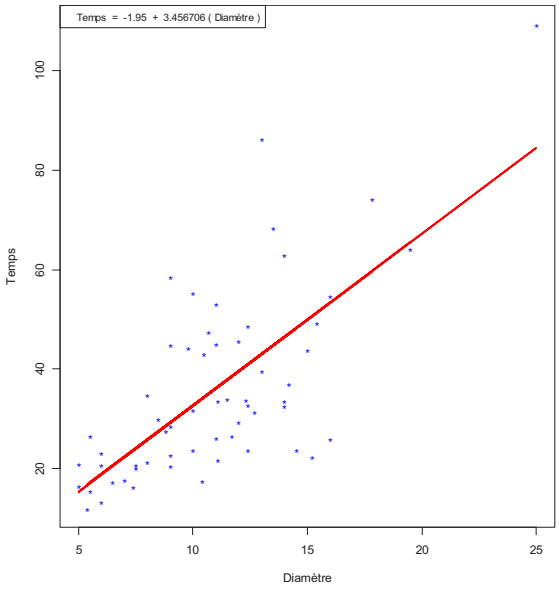
f) $b_1 - 2\hat{\sigma}_{b_1} \leq \beta_1 \leq b_1 + 2\hat{\sigma}_{b_1} \Leftrightarrow 1{,}323 \leq \beta_1 \leq 1{,}882$

g) $r = 0{,}8562459$; $Z = 11{,}60296$.

h) $\hat{\mu} = \hat{\mu}_{y.15} = 3{,}609766$. Intervalle de confiance : $3{,}494421 \leq \mu_{y.15} \leq 3{,}725112$. Il s'agit d'un intervalle de confiance non pas pour la distance moyenne, mais plutôt pour la moyenne du *logarithme* de la distance. $e^{3{,}494421} \leq$ Moyenne des logarithmes $\leq e^{3{,}725112} \Leftrightarrow 32{,}93 \leq$ Moyenne des logarithmes $\leq 41{,}47$. (Voir la remarque à la fin du numéro 12.6.)

Problème 12.13

a)



b) $b_1 = 3{,}456706$; $b_o = -1{,}954716$. Droite de régression : Temps = -1,95 + 3,46(Diamètre)

c) $r = 0{,}700$

d) $\hat{\sigma}_{y.x} = = 13{,}69307$

e) $\hat{\sigma}_{b_1} = 0{,}4666758$

f) $b_1 - 2\hat{\sigma}_{b_1} \leq \beta_1 \leq b_1 + 2\hat{\sigma}_{b_1} \Leftrightarrow 2{,}523354 \leq \beta_1 \leq 4{,}390057$

g) $r = 0{,}7003261$; $Z = 7{,}471774$

h) $\hat{\mu} = 49{,}89587$; $\hat{\sigma}_{\hat{\mu}} = 2{,}605895$; Intervalle de confiance : $44{,}68408 \leq \mu_{y.15} \leq 55{,}10766$

Problème 12.14

a)

Objet	b_1	$\hat{\sigma}_{b_1}$	$b_1/\hat{\sigma}_{b_1}$
Assiettes	4,551	1,764	2,580
Casseroles	0,595	2,996	0,199
Plateaux	4,439	0,590	7,523
Plats de service	0,625	0,386	1,618
Bols	2,154	0,707	3,047

b) Différence entre plateau et plats de service:

Différence de pentes :b_1-d_1 = 3,814169; écart-type de la différence : $\sqrt{\hat{\sigma}_{b_1}^2+\hat{\sigma}_{d_1}^2}$ = 0,7051811; $Z=\dfrac{b_1-d_1}{\sqrt{\hat{\sigma}_{b_1}^2+\hat{\sigma}_{d_1}^2}}=5{,}4.$

Différence entre assiettes et plats de service

Différence de pentes :b_1-d_1 = 3,926211 ; écart-type de la différence : $\sqrt{\hat{\sigma}_{b_1}^2+\hat{\sigma}_{d_1}^2}$ = 1,80598 ; $Z=\dfrac{b_1-d_1}{\sqrt{\hat{\sigma}_{b_1}^2+\hat{\sigma}_{d_1}^2}}=2{,}17.$

Problème 12.15

a) b_1 = 0,8344424; b_o = 707.8921

b) r = 0,7285098

c) $\hat{\sigma}_{y.x}$ = 4548,282

e) $\hat{\sigma}_{b_1}$ = 0,1132539

f) $b_1 - 2\,\hat{\sigma}_{b_1} \le \beta_1 \le b_1 + 2\,\hat{\sigma}_{b_1} \Leftrightarrow$ 0,6079345 $\le \beta_1 \le$ 1,06095

g) Z = 7,367888

h) $\hat{\mu}$ = 21568,95

i) $\hat{\sigma}_{\hat{\mu}}$ = 655,2347; $\hat{\mu} - 2\,\hat{\sigma}_{\hat{\mu}}$ = 20258,48; $\hat{\mu} + 2\,\hat{\sigma}_{\hat{\mu}}$ = 22879,42

Problème 12.16

a) L'intervalle calculé au dernier numéro était $[b_1\text{-}2\,\hat{\sigma}_{b_1}\ ;\ b_1\text{+}2\,\hat{\sigma}_{b_1}]$ = [0,6079345 ; 1,0609503]. Cet intervalle contient, en effet β_1, un événement qui est censé se produire environ 95 % des fois.

b) s_x = 5737,147; $\sigma_{b_1} = \dfrac{\sigma_{y.x}}{\sqrt{\sum(x_i-\bar{x})^2}}$ = 0,1061314

c) L'intervalle de confiance est $[b_1\text{-}2\,\sigma_{b_1}\ ;\ b_1\text{+}2\,\sigma_{b_1}]$ = [0,6221796 ; 1,046705]

Erratum : Remplacer $\mu_{y.50000}$ par $\mu_{y.25000}$

d) $\mu_{y.25000} = \beta_o + \beta_1(25000)$ = 20672,87

e) L'intervalle de confiance calculé au numéro précédent est [20258,48; 22879,42]. Il contient donc $\mu_{y.25000}$ = 20672,87 comme il le devrait environ 95 % des fois.

f) L'écart-type de $\hat{\mu}_{y.25000}$ n'a pas à être estimé. Il est connu et se calcule par la formule $\sigma_{\hat{\mu}} = \sigma_{y.x}\sqrt{\dfrac{1}{n}+\dfrac{(25000-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$ = 614,027.

g) $\hat{\mu}$ = 21568.95. Intervalle de confiance $\hat{\mu} - 2\,\sigma_{\hat{\mu}} \le \mu_{y.25000} \le \hat{\mu} + 2\,\sigma_{\hat{\mu}} \Leftrightarrow 20340.9 \le \mu_{y.25000} \le 22797.01$

h) $\dfrac{b_o-\beta_o}{\sigma_{b_o}} = 0{,}3258$

Problème 12.17

	Femmes	Hommes
b_1 , g_1	4	2
b_o , g_o	2	27
$\hat{\sigma}_{y.x}$	6,020236	10,458710
$\hat{\sigma}_{b_1}$, $\hat{\sigma}_{g_1}$	0,2465985	0,3922323

a) *Voir le tableau*

b) *Voir le tableau*

c) $\hat{\sigma}_{b_1-g_1} = \sqrt{\sigma_{b_1}^2+\sigma_{g_1}^2}$ = 0,4633109

d) $\dfrac{b_1 - g_1}{\sqrt{\hat{\sigma}^2_{b_1} + \hat{\sigma}^2_{g_1}}} = 4{,}316756$

e) $\mu_{fy.13} = 54$; $\hat{\mu}_{hy.13} = 54\ 53$

f) La différence est de 1, les femmes ayant le score le plus élevé.

Erratum : Remplacer la question g) par

g) Estimer l'écart-type de l'estimateur en f).

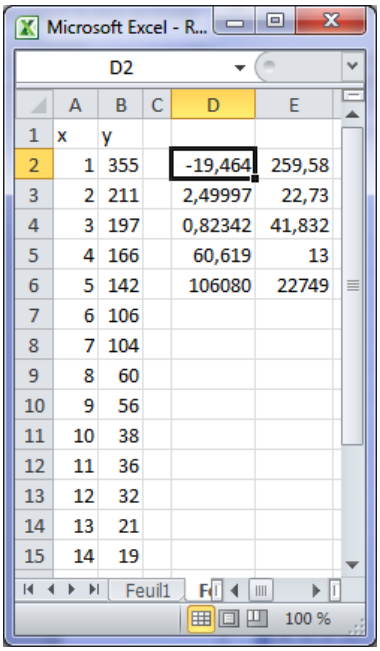
g) $\hat{\sigma}_{\hat{\mu}_{fy.13}} = 0.5499386$; $\hat{\sigma}_{\hat{\mu}_{fy.13}} = 1{,}2333507$; $\hat{\sigma}_{\hat{\mu}_{fy.13} - \hat{\mu}_{hy.13}} = \sqrt{\hat{\sigma}^2_{fy.13} + \hat{\sigma}^2_{fy.13}} = 0{,}7405201$

h) Non, car $Z = \dfrac{\hat{\mu}_{fy.13} - \hat{\mu}_{hy.13}}{\hat{\sigma}_{\hat{\mu}_{fy.13} - \hat{\mu}_{hy.13}}} = 1{,}35.$

Commandes Excel

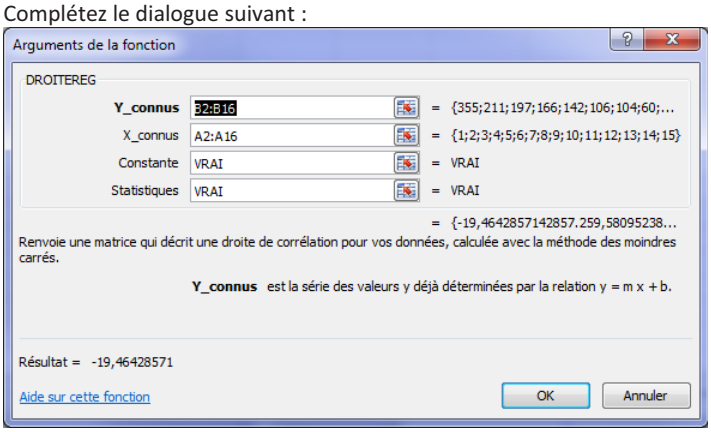
Exemple de calcul avec Excel. Illustré avec les données du numéro 12.6.

Les données se trouvent dans les colonnes `A` et `B`, lignes `2` à `16`.
À partir d'un point quelconque, sélectionnez une plage de 5 lignes et 2 colonnes. Ici, nous avons choisi la plage `D2` à `E6`.



	A	B	C	D	E
1	x	y			
2	1	355		-19,464	259,58
3	2	211		2,49997	22,73
4	3	197		0,82342	41,832
5	4	166		60,619	13
6	5	142		106080	22749
7	6	106			
8	7	104			
9	8	60			
10	9	56			
11	10	38			
12	11	36			
13	12	32			
14	13	21			
15	14	19			

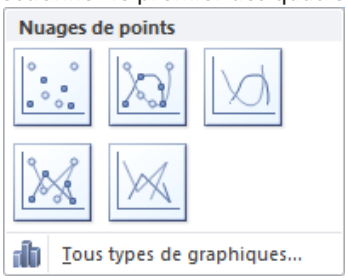
Dans l'onglet **Formules** cliquez sur **Insérer une fonction** puis choisir **DROITEREG**. Complétez le dialogue suivant :



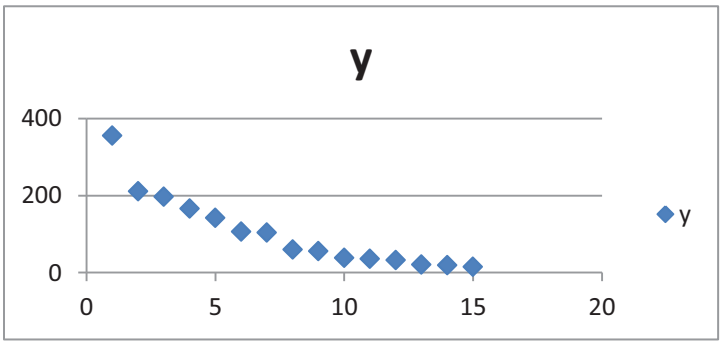
Les deux premières cases identifient les valeurs de *y* et celles des *x*, respectivement. Inscrire VRAI aux deux dernières cases. Cela signifie que vous voulez obtenir toutes les données qu'Excel peut fournir.
Ne cliquez pas sur OK. Faites plutôt
`Ctrl+Majuscule+Entrée`
Vous obtiendrez alors, dans la plage `D2` à `E6`, les statistiques suivantes :

b_1	b_o
$\hat{\sigma}_{b_1}$	$\hat{\sigma}_{b_o}$
r^2	$\hat{\sigma}_{y.x}$
Z^2	n-2
$b_1^2\sum(x_i-\bar{x})^2$	$(n-2)\hat{\sigma}_{y.x}^2$

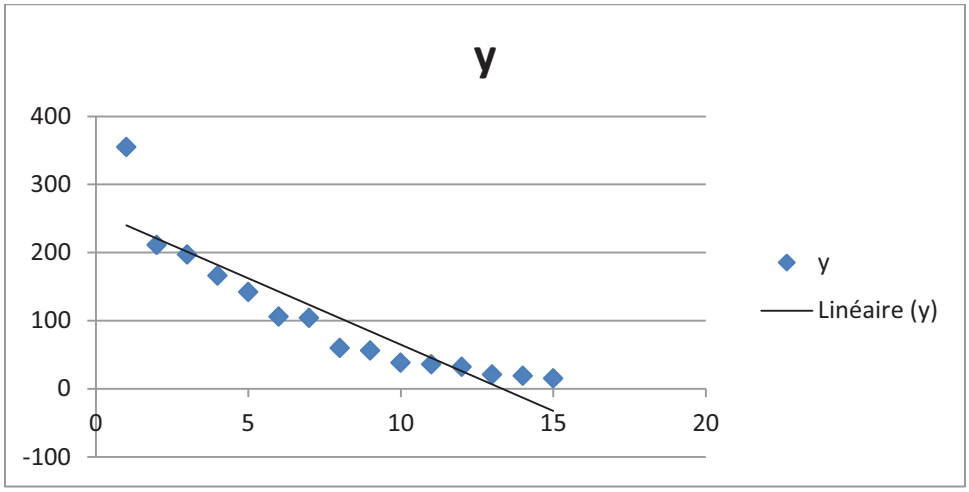
Sélectionner la plage contenant les données (y compris les titres). Il est préférable de placer la colonne des *x* à gauche.
Faites `Insertion > Nuage` > puis sélectionnez le premier des quatre choix suivants :



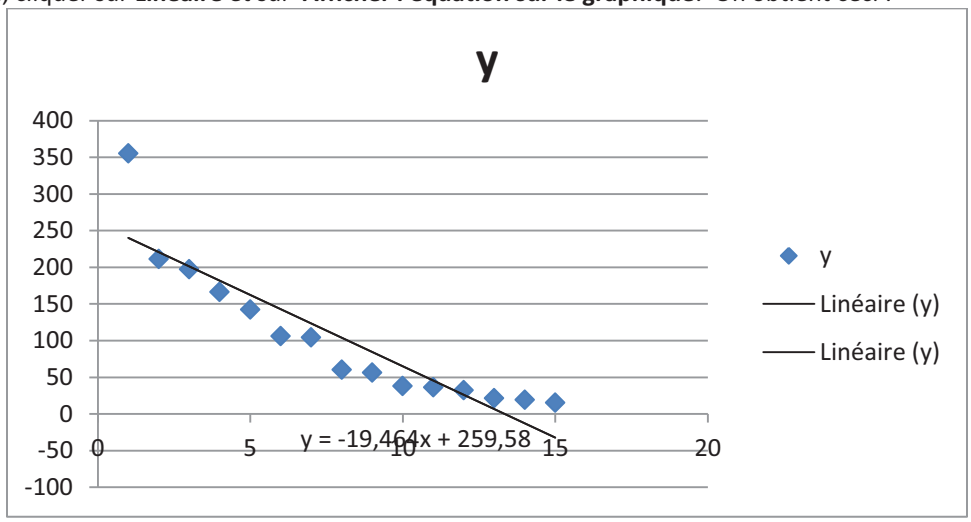
Vous obtiendrez le graphique suivant (que vous pourrez ensuite mettre en forme à l'aide des diverses fonctionnalités d'Excel)



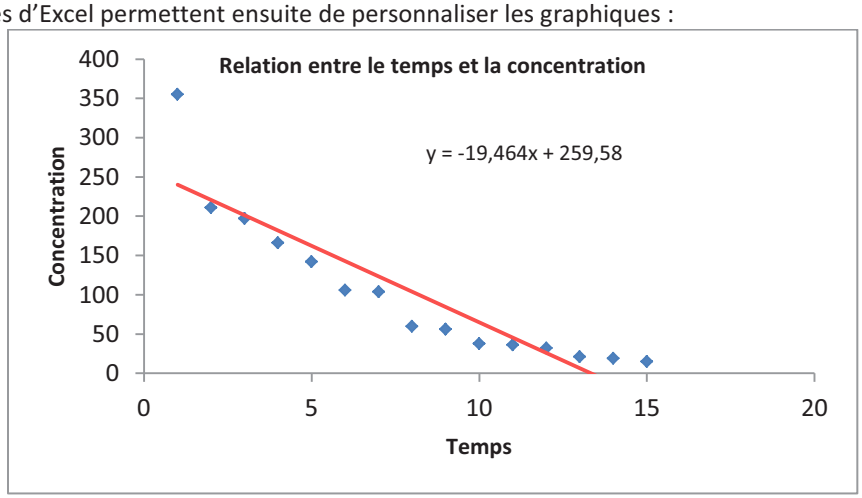
Pour obtenir la droite de régression : Sous l'onglet **Disposition**, dans le groupe **Analyse**, cliquez sur **Courbe de tendance**, puis cliquez sur **Courbe de tendance linéaire**.
Vous verrez ceci :



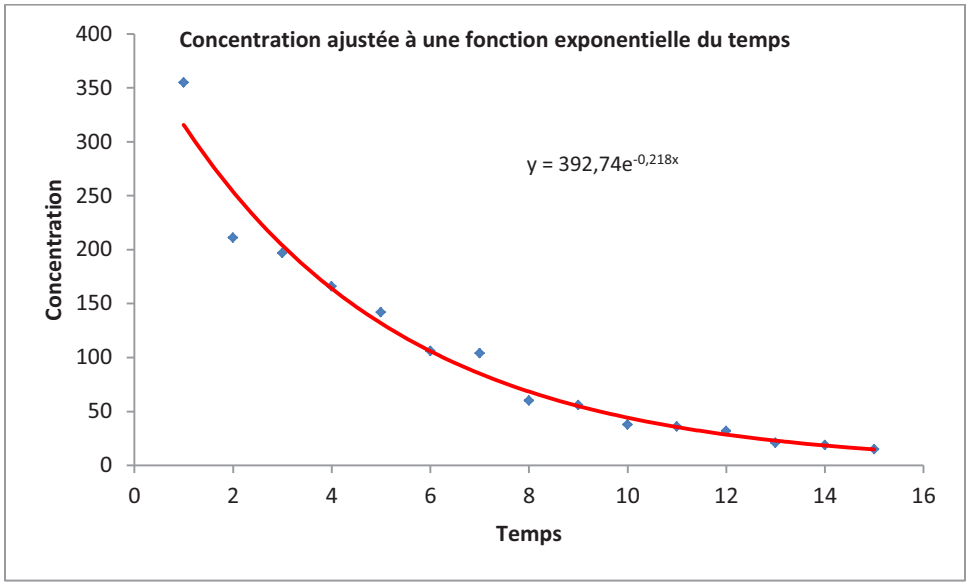
Inscrire l'équation sur le graphique : Sous l'onglet **Disposition**>**Courbe de tendance**>**Autres options de la courbe de tendance**, cliquer sur **Linéaire** et sur **Afficher l'équation sur le graphique**. On obtient ceci :



Les fonctionnalités d'Excel permettent ensuite de personnaliser les graphiques :



Pour un ajustement exponentiel : **Disposition**>**Courbe de tendance**>**Autres options de la courbe de tendance**.
Cliquez sur **Exponentielle** et sur **Afficher l'équation sur le graphique**. Après un certain nettoyage, on obtient ceci :


Concentration ajustée à une fonction exponentielle du temps
y = 392,74e-0,218x
Concentration
Temps
400
350
300
250
200
150
100
50
0
0
2
4
6
8
10
12
14
16