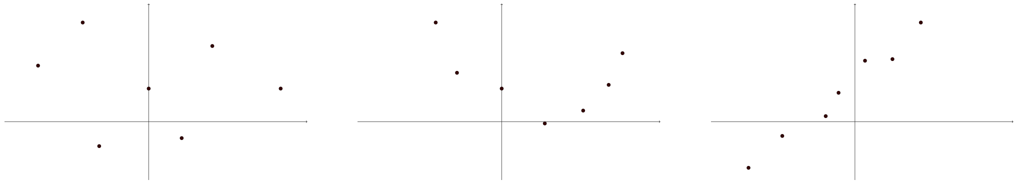
TD2 : Régression linéaire

Objectif : on cherche à établir un ajustement linéaire entre différentes quantités pour prédire le comportement futur de nos données.

I. Introduction

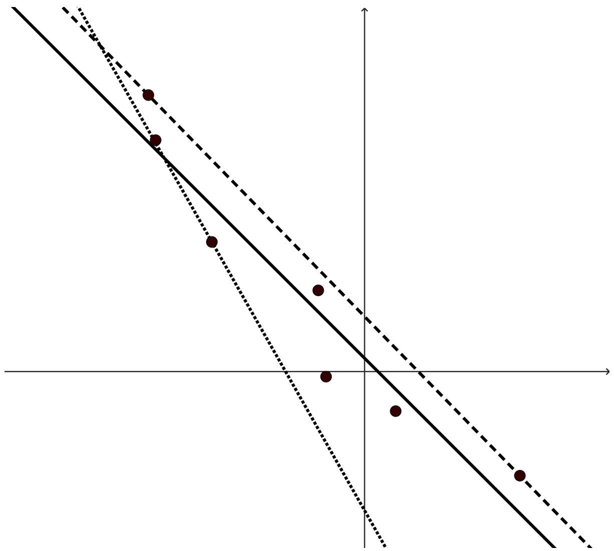
Exercice 1

Quels sont les nuages de points qui semblent pouvoir être correctement approchés par une droite ?



Exercice 2

Parmi les différentes droites proposées pour approcher le nuage de point suivant, quelle est celle qui semble la plus pertinente ?



II. Révision : statistiques

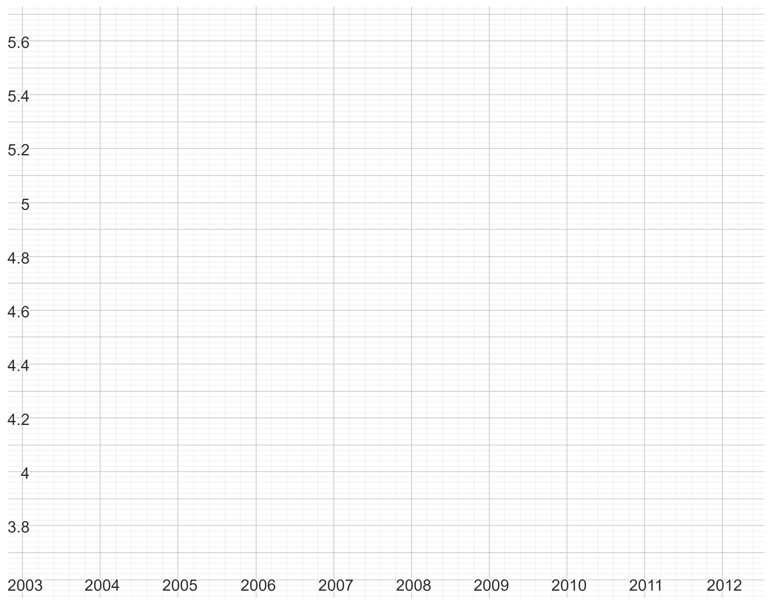
Exercice 3

Dans une entreprise de construction automobile, on donne le temps t de montage en heures d'un véhicule en fonction de l'année x.

On arrondira les résultats au millième.

x	2005	2006	2007	2008	2009	2010	2011
t	26.2	23.7	21.4	18.5	16.8	15.4	14.6
$y = \sqrt{t}$							

1. Quel est l'effectif total n_{tot} de nos données ? $n_{tot} =$
2. On pose $y = \sqrt{t}$. Remplir la troisième ligne du tableau.
3. Sur le graphique, tracer y en fonction de x en utilisant un **nuage de points**.



Exercice 4

1. La **moyenne** est un **indicateur de position**.

   Elle permet de "résumer" une série en une seule valeur.

   Calculer les moyennes suivantes :

   $\overline{x} =$

   $\overline{x^2} =$

   $\overline{y} =$

   $\overline{y^2} =$

   $\overline{xy} =$

2. La **variance** est la différence entre la moyenne des carrés $\overline{x^2}$ et le carré de la moyenne $(\overline{x})^2$.

   La variance est **toujours positive**.

   Variance des x : $\boxed{var(x) = \overline{x^2} - (\overline{x})^2}$

   Calculer les variances suivantes :

   $var(x) =$

   $var(y) =$

3. L'**écart-type** est **indicateur de dispersion**.

   Il donne une "distance" entre la moyenne et la série.

   Quand l'écart-type est grand, les données sont très éparpillées autour de la moyenne.

   Quand l'écart-type est petit, les données sont reserrées autour de la moyenne.

   Ecart-type des x : $\boxed{\sigma_x = \sqrt{var(x)}}$

   $\boxed{var(x) = \sigma_x^2}$

   Calculer les écarts-types suivants :

   $\sigma_x =$

   $\sigma_y =$

4. La **covariance** $cov(x, y)$ est une quantité qui illustre le comportement des x par rapport aux y.

   Si y est croissant quand x est croissant, la covariance est positive.

   Si y est décroissant quand x est croissant, la covariance est négative.

   Covariance des x et y : $\boxed{cov(x, y) = \overline{xy} - \overline{x} \times \overline{y}}$

   Calculer $cov(x, y) =$

III. Révision : régression linéaire simple

Exercice 5

1. Le **coefficient de corrélation linéaire** R permet d'évaluer si il est pertinent de modéliser notre nuage de points par une droite.

   On a toujours $-1 < R < 1$.

   Si y est croissant quand x est croissant, $R > 0$.

   Si y est décroissant quand x est croissant, $R < 0$.

   Le **coefficient de détermination** est R^2. C'est ce coefficient qui est calculé par Excel lors d'un ajustement linéaire.

On a toujours $0 < R^2 < 1$.

Plus R^2 est proche de 1, plus les points de notre nuages sont alignés, et donc plus il est pertinent faire une modélisation linéaire.

$\boxed{R = \frac{cov(x,y)}{\sigma_x \sigma_y}}$

Calculer R et R^2

$R =$

$R^2 =$

2. Est-ce qu'il vous semble pertinent de tracer une droite D d'équation $y = ax + b$ pour modéliser vos données d'après vos calculs ?

Exercice 6 Droite d'ajustement linéaire

1. Calculer le coefficient directeur a de la meilleure droite D qui modélise le nuage.

   $\boxed{a = \frac{cov(x,y)}{var(x)}}$

   $a =$

2. Calculer l'ordonnée à l'origine b de la meilleure droite D qui modélise le nuage.

   $\boxed{b = \overline{y} - a \times \overline{x}}$

   $b =$

3. Donner l'équation de la droite D

   $\boxed{D \ : \ y = ax + b}$

   $D \ : \ y =$

4. On considère deux points de la droite D :
   - le point A, d'abscisse $x_A = 2003$
   - le point B d'abscisse $x_B = 2012$.

   Calculer les ordonnées de ces points
   - $y_A =$
   - $y_B =$

   Donc $A(2003\ ;\quad)$ et $B(2012\ ;\quad)$.

5. La droite D est la même que la droite (AB).

   Tracer la droite D sur le même graphique que le nuage de points.

6. Placer sur le graphique le **point moyen G** de coordonnées $(\bar{x}, \bar{y})$

Exercice 7 Prédictions

On pose dorénavant $\boxed{\hat{y} = a \times x + b}$

Répondre en utilisant le graphique quand c'est possible.

1. a) Quelle valeur $\hat{y}$ peut-on prédire à l'aide de notre modèle pour $x = 2004$?
   b) À quelle valeur t cela correspond-t-il ?
   c) Interpréter le résultat dans le contexte de l'exercice.

2. a) On cherche à résoudre $\hat{y} < 5$. À quelles valeurs de t cela correspond-t-il ?
   b) Résoudre l'inéquation.
   c) Interpréter le résultat dans le contexte de l'exercice.

3. a) Quelle valeur $\hat{y}$ peut-on prédire à l'aide de notre modèle pour $x = 2025$?
   b) À quelle valeur t cela correspond-t-il ?

c) Interpréter le résultat dans le contexte de l'exercice.

4. a) On cherche à résoudre $\hat{y} < 2$. À quelles valeurs de t cela correspond-t-il ?
   b) Résoudre l'inéquation.
   c) Interpréter le résultat dans le contexte de l'exercice.

Exercice 8 **Erreur-type et résidus**

Répondre à cet exercice en utilisant Excel.

1. Dans un nouveau fichier Excel, remplir les cinq premières lignes du tableau suivant

x							
y							
$\hat{y}$							
$y-\hat{y}$							
$(y-\hat{y})^2$							
u							

2. On définit l'**erreur-type** s comme :

   $s = \frac{1}{\sqrt{n_{tot}-2}} \times \sqrt{(y_1-\hat{y}_1)^2 + ... + (y_7-\hat{y}_7)^2}$

   Calculer l'erreur-type grâce à Excel :

   $s =$

3. On définit les **résidus** : $\boxed{u = \frac{y-\hat{y}}{s}}$

   Calculer les résidus dans la dernière ligne du tableau d'Excel.

4. À l'aide d'Excel, tracer les résidus u en fonction des x .

IV. Généralisation : régression linéaire multiple

On change de série statistique pour ce dernier exercice.
Les données sont dans le fichier Excel du TD2, onglet Exercice 9

Exercice 9

On cherche à modéliser linéairement la **valeur immobilière** d'un immeuble en fonction de sa **superficie utile**, de son **nombre de bureaux**, de son **nombre d'entrées** et de son **âge**.

On cherche des nombres réels A_0, A_1, A_2, A_3 et A_4 tels que :

$$\textbf{valeur} = A_0 + A_1 \times \textbf{superficie} + A_2 \times \textbf{nb bureaux} + A_3 \times \textbf{nb entrées} + A_4 \times \textbf{âge}$$

Merci de bien suivre la procédure décrite ci-dessous pour installer l'add-in *Analysis ToolPak* dans Excel :

- Cliquer sur *Fichier/Options/Compléments*
- Cliquer sur *Analysis ToolPak* puis sur *Atteindre* (tout en bas)
- Cocher *Analysis ToolPak*, puis cliquer sur *OK*
- Aller dans *Utilitaire d'analyse* (à droite de votre barre d'outils)
- Sélectionner les données qui vous intéressent
- Analyser à l'aide de *Régression linéaire*
- Tester plusieurs possibilités et retenir celle qui vous semble la meilleure.

Quelques précisions :

- Quand le p-value est petit, la variable considérée a un pouvoir prédictif intéressant.
- La statistique F permet de déterminer si les résultats présentant une valeur de R^2 élevée sont le fruit du hasard.

  Idéalement, on cherche à avoir F grand.