1. *Liaison significative*

a. On interprète la phrase "Fumer double le risque de maladie cardio-vasculaire" par la relation entre probabilités conditionnelles $\mathrm{P}(M|F) = 2\mathrm{P}(M|\text{non } F)$ où M est l'évènement "souffrir un jour d'une maladie cardio-vasculaire" et F l'évènement "être un fumeur régulier". On suppose $\mathrm{P}(M|F) = 0.6$. Calculer les coefficients de liaison entre M et F et entre M et non F dans chacune des situations suivantes : 90% de la population est fumeur ; 10% de la population est fumeur ; 50% de la population est fumeur.

Pour chacune des situations ci-dessus, la liaison éventuellement observée entre M et F ou entre M et non F est elle significative pour M ? (Une liaison entre deux évènements E et F est significative pour E si sachant E l'évènement F n'est pas rare, concrètement (pour le cours) si $\mathrm{f}_{F|E} \geq 0.1$).

2. Un calcul de Chi-deux.

Une population de 410 individus est étudiée via deux caractères qualitatifs X et Y prenant pour valeurs x_1, x_2, x_3 et y_1, y_2, y_3, y_4 respectivement, et dont les effectifs conjoints sont donnés par le tableau ci-dessous :

X \ Y	y_1	y_2	y_3	y_4
x_1	63	42	0	0
x_2	0	0	0	195
x_3	0	0	110	0

Que vaut $\chi^2(X, Y)$? Commentaires ?

3. *Coefficient de corrélation*

On a mesuré la longévité des piles pour un ensemble de 40 piles de marque A, 30 piles de marque B et 50 piles de marque C. La longévité moyenne des piles de marque A est de 71.8 (minutes) ; celle de marque B est 60.2 ; celle de marque C est 75.3.

Quelle est la longévité moyenne des piles testées ?

L'écart type des longévités dans la population entière est de 7. Que vaut le coefficient de corrélation de la longévité selon la marque ? Commentaires ?

4. Une société fait une analyse de ses coûts de publicité (en million d'euros) et du nombre d'unités vendues (en million) d'un produit, chaque mois pendant douze mois. Elle obtient les données suivantes (publicité mensuelle, ventes mensuelles) :

```
(5.2, 1.3), (5.2, 2.1), (11.5, 2.4), (5.1, 1.7), (13.5, 3.1), (11, 2.8), (8.6, 2.7), (17.5, 4.2),
(18.5, 3.2), (12.4, 3.7), (11, 2.6), (11, 3)
```

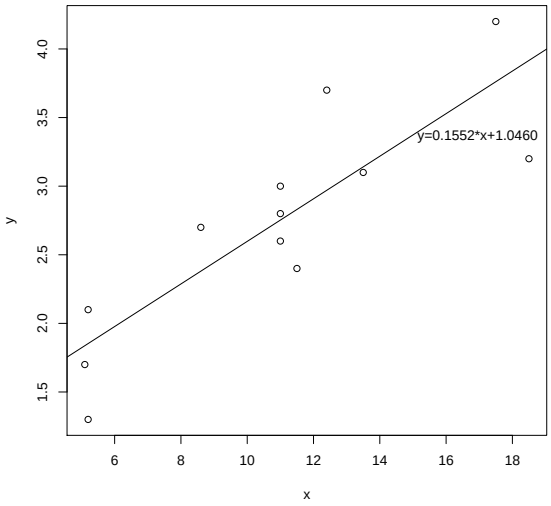
a. Calculer la moyenne des coûts de publicité mensuels et des ventes mensuelles. Calculer les variances associées, puis la covariance des coûts et ventes.

b. Dessiner le nuage de points et marquer le centre de gravité du nuage.

c. Calculer les coefficients de la régression linéaire et dessiner la droite de régression

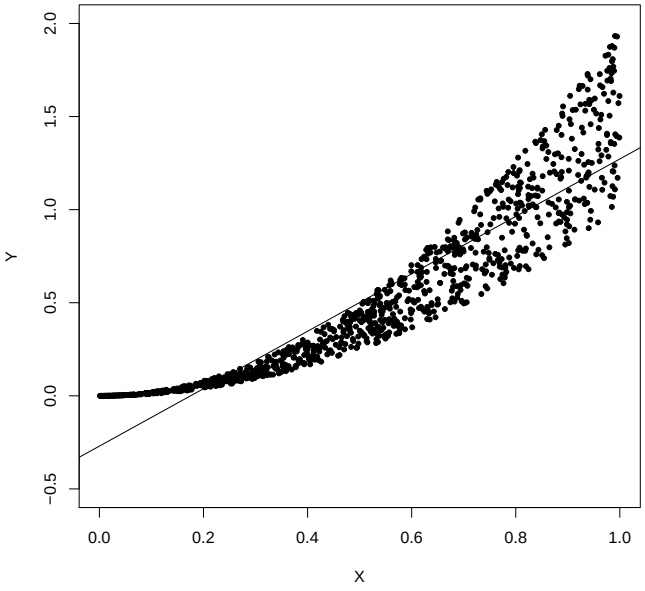
d. Observe t-on l'indépendance des deux caractères ?

e. Supposons que le bénéfice par unité de produit vendu soit de 5 euros. En ce basant sur la régression linéaire, la société aura t-elle plutôt intérêt de faire plus de publicité ou au contraire moins de publicité sur le produit ?



Solution partielle :

5. Une population de 1000 individus est étudiée à travers deux caractères quantitatifs X et Y. On représente ci-dessous le nuage des points $(X(i), Y(i))$ et la droite de régression $y = ax + b$. Pour chaque individu i on note $R(i)$ la différence entre la valeur de Y et la valeur prédite par X et la régression linéaire : R(i)=Y(i)-aX(i)-b.



a. D'après le graphique, que valent approximativement les coefficients a et b de la droite de régression ?

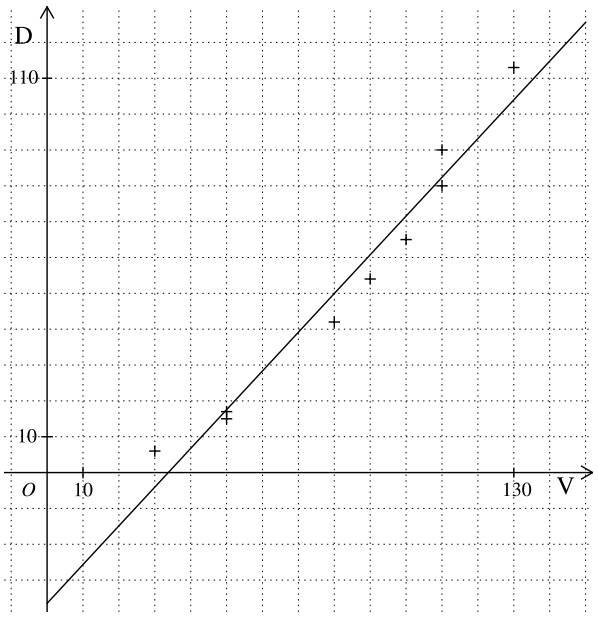
b. Sachant que la moyenne de X vaut $\frac{1}{2}$, que vaut approximativement la moyenne de Y ?

c. Que vaut la moyenne de Y conditionnée à l'évènement "X est proche de 0" ? Que vaut la moyenne de R conditionnée à "X est proche de 0" ?

d. Que vaut les moyennes de Y et de R conditionnées à "X est proche de 1" ?

e. La régression linéaire est elle satisfaisante ? Quel changement de variable sur X peut on essayer pour améliorer la régression ?

6. (Examen de 2011) Un constructeur automobile étudie les distances de freinage d'un véhicule à différentes vitesses. On représente ci-dessous le nuage des points $(V(i), D(i))$ où $V(i)$ est la vitesse (en km/h) avant freinage et $D(i)$ la distance de freinage (en m) lors de la i-ième expérience. On fait l'hypothèse qu'il n'y a pas de points confondus. On représente également la droite de régression associée aux données.



a. Quelle est la taille de la population ? Quelle est l'étendue du caractère V ? et de D ? Quelle est très approximativement la moyenne de V ?

b. Soit E l'évènement "la vitesse est supérieure à 85". Soit F l'évènement "la distance de freinage est supérieure à 60".

Calculer les fréquences des évènements E et F puis la fréquence conditionnelle de F sachant E. L'évènement F est il pratiquement indépendant de E ? Justifier.

c. Le caractère D est il indépendant de V ? Justifiez (plusieurs justifications sont possibles).

d. Que peut on dire du centre du nuage de points par rapport à la droite de régression ? On a calculé la moyenne des distances de freinage et trouvé 53.5. Que vaut approximativement la moyenne des vitesses ?

e. Quelle est la distance de freinage prédite par la régression linéaire pour $V = 10$? Cette valeur est elle pertinente ? Qu'en conclut on quant au modèle donné par la régression linéaire ?