Questions supplémentaires portant sur les tests

1- Le tableau suivant présente les fréquences observées d'un échantillon et les fréquences théoriques d'une certaine loi de distribution entièrement spécifiée (i.e. aucun paramètre n'a eu à être estimé). Faites le test χ^2 permettant de vérifier si la distribution proposée est adéquate (niveau 5%).

Intervalle	[0, 1[	[1, 2[	[2, 3[	[3, ∞ [	Total
f observé	10	55	30	5	100
f théorique	9	50,4	34,5	6,1	100

2- Dans un sondage touchant 1000 personnes, 645 ont dit préférer l'option A aux autres options.
a) Peut-on conclure que la proportion "p" est significativement supérieure à 0.6?
b) Quelle est la plus petite valeur de α qui permettrait de rejeter H_0 ?

3- Un échantillon aléatoire de 9 cigarettes d'un paquet montre un taux moyen de nicotine de 4.2g par cigarette. Le paquet mentionne un taux moyen de 3.5.
a) Peut-on mettre en doute l'indication sur le paquet sachant que le taux de nicotine est N(μ,1.96)?
b) Quelle est la probabilité de conclure que le fabricant a tort (i.e. rejeter l'affirmation sur le paquet) si la vraie moyenne de nicotine est 3,8?
c) Combien de cigarettes devrait on examiner pour assurer une probabilité de 0,9 en b) ?

4- Le tableau suivant montre les résistances en compression de deux bétons. Les unités sont des MPa/10.

Béton 1	295	319	304	302
Béton 2	318	316	312	318

a) Faites le test pour vérifier si les variances sont égales (niveau 5%)
b) Faites le test approprié pour vérifier si la moyenne du 2^e béton est effectivement supérieure à celle du 1er (niveau 5%)

5- Soit X distribué suivant une loi uniforme sur [0,c]. On veut tester H_0 : c=1 vs H_1 c>1. On recueille une seule observation et on rejette H_0 si x_1>0,9.
a) Quel est le risque α de ce test?
b) Quelle est la puissance du test si en réalité c=1,5?

6- On s'intéresse à la durée de vie des pneus de marque A (en dizaine de milliers de kilomètres). On désire tester H_0 : μ=50 vs H_1 μ<50 au niveau 5% . On suppose X~N(μ , 25)
a) Quelle est la puissance du test si n=9 et μ=45?
b) Quelle valeur de « n » assure une puissance >0,9?

7- Le tableau suivant montre la répartition du succès scolaire de 150 enfants en fonction du revenu de leurs parents. Ces deux variables sont-elles indépendantes (niveau 5%)?

	Revenu			
Succès	Faible	Moyen	Élevé	Total
Faible	10	20	10	40
Moyen	10	40	10	60
Élevé	10	30	10	50
Total	30	90	30	150

8- On veut tester si une distribution binomiale Bin(3,1/2) s'ajuste aux données suivantes (niveau 5%) :

x	0	1	2	3
Observé	7	12	18	3
Bin(3,1/2)	5	15	15	5

9- Un manufacturier de caoutchouc synthétique affirme que son produit présente une dureté Shore de 64,3. L'écart-type est de 2.

a) Combien prendre d'échantillons pour assure un test de niveau 5% et de puissance 90% lorsque la vraie dureté est 64.3 ± 1.

b) Si $\bar{x}$ =65 et n=43, quelle est la conclusion du test H_0 μ =64.3 vs H_1 $\mu \neq 64.3$?

c) Quelle est l'erreur de seconde espèce (β) si n=43 et la vraie dureté Shore est 65?

d) Combien d'observations prélever si l'on veut rejeter H_0 avec un risque de 5% lorsque $\bar{x}$ =65 (avec un test bilatéral)?

10- Deux lignes de fabrication de ciment ont fourni les concentrations suivantes en C_3S (les valeurs sont multipliées par 2 pour simplifier les calculs).

Ligne 1	140	135	140	138	135	138	140
Ligne 2	135	138	136	140	138	135	139

Les deux lignes ont-elles même variance (niveau 10% bilatéral)?

11- Des tiges métalliques sont achetées de deux entreprises différentes. On soumet 10 tiges de chaque entreprise à un test de tension. Les moyennes et variances obtenues pour la résistance en tension sont :

$\bar{x}_1$ =50,29 et $\bar{x}_2$ =55,72 ; s_1^2=22,72 et s_2^2=14,91.

Peut-on conclure que les entreprises fournissent des tiges de même résistance? (Faites le test d'égalité des variances (bilatéral 10%) avant celui des moyennes (bilatéral 5%)).

Questions supplémentaires sur les régressions

Note :

SCT : somme des carrés de Y,
SCT_m : somme des carrés de (Y-m), m : moyenne de Y (ou Y bar)
SCR : somme des carrés de Y_p (Y prédit ou chapeau)
SCR_m : somme des carrés de (Y_p-m)
SCE : somme des carrés des résidus (e)
SCM : somme des carrés dûs à la moyenne (n m^2)
CMx : carrés moyens (i.e. SCx divisée par ses degrés de liberté)

1- Vous effectuez un stage au concentrateur d'une mine de Zn et vous êtes en charge du calibrage d'un analyseur en continu aux rayons X du concentré, le Courier-300. Pour effectuer ce calibrage, un échantillon est analysé à intervalles réguliers par l'analyseur Courier-300 et par analyse chimique conventionnelle. La teneur obtenue à l'analyse chimique est considérée comme la vraie teneur de l'échantillon. Le but est de fournir une équation continuellement mise à jour permettant de corriger la teneur obtenue au Courier-300 de façon à mieux estimer les vraies teneurs et permettre ainsi un meilleur ajustement des procédés de flottation.

On vous fournit les quantités suivantes :

Nombre d'échantillons: 5

	Courier-300	Analyse chimique
Somme des valeurs	241	260
Somme des carrés des valeurs	11700	13618
Somme des produits croisés		12617

a) Un échantillon de concentré analysé au Courier-300 montre une teneur de 50% Zn. Quelle serait la teneur corrigée? (Utilisez un modèle avec constante)

b) Est-ce que le modèle de régression établi en a) est significatif? Faites le test requis.(Aide : vous pouvez soit i) calculer le coefficient de corrélation simple,en déduire le R^2,…ou ii) utiliser le fait que $Y'Y_p=Y'Xb=Y_p'Y_p=SCR$)

2- Le tableau suivant présente diverses sommes des carrées pour la régression (modèle avec constante) de données comportant 22 observations et 3 variables. La colonne de gauche est obtenue avec l'ensemble des données, la colonne de droite est obtenue après avoir retirée la donnée #9 et en utilisant les mêmes variables. La somme des carrés des différences entre les valeurs prédites par les deux modèles vaut 4.

Somme des carrés	Modèle avec toutes les observations	Modèle sans l'observation #9
SCTm	503	457
SCRm	254	231
SCE	249	226

Quelle est l'influence de l'observation #9 sur la régression? Qu'en concluez-vous?

3- On mesure une teneur sur 2 sites différents. On a prélevé n1 observations sur le 1er site et n2 sur le second.

Expliquez comment vous pourriez utiliser un programme de régression pour tester l'égalité des moyennes des teneurs des 2 sites.

4- On effectue des relevés de diagraphies en forage à l'aide de 2 sondes pour prédire la porosité (exprimée en %). Celle-ci a été déterminée pour 39 carottes prises le long du forage . La 1ère sonde mesure 3 signaux et la régression procure un R^2 de 0.8. La 2e sonde mesure 2 signaux et la régression procure un R^2 de 0.7. Lorsque les 5 signaux des 2 sondes sont utilisés conjointement dans la régression, le R^2 atteint 0.85.

D'un point de vue statistique, vaut-il la peine d'utiliser les 2 sondes? La variance expérimentale de la porosité (s_y^2) atteint 112%^{2}.

5- Dans un modèle de régression linéaire $Y=b_0+b_1X+e$, on sait que $b_0=\overline{Y}-b_1\overline{X}$.

Utilisez ce fait pour modifier le modèle initial et montrer ainsi que b_1 *peut être obtenu directement d'une régression sans constante.*

6- Pour un aquifère à nappe captive sans recharge et soumis à un pompage, on atteint un état d'équilibre liant le rabattement (s) (i.e. diminution de la charge par rapport à l'état initial) à la distance au puits de pompage (r). On a alors: $s=b_0+b_1\ln(r)+e$. Le coefficient b_1 est lié à la transmissivité par: $b_1=\frac{-Q}{2\pi T}$ où Q est le débit pompé (100cm^3/s) et T est la transmissivité. On a disposé 3 piézomètres selon une direction en s'éloignant du puits et l'on a observé:

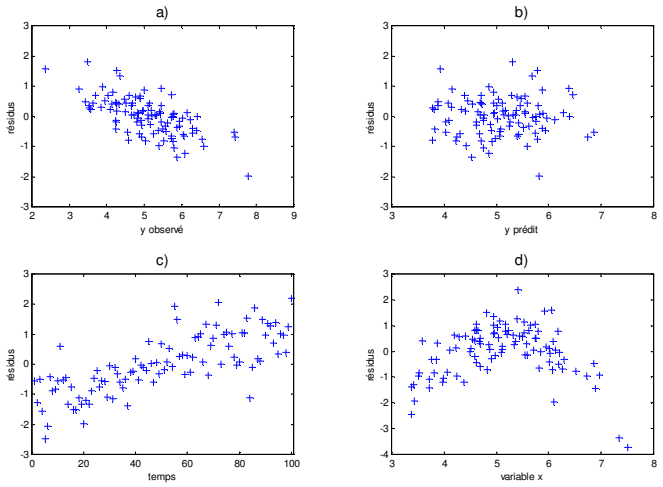
Piézomètre #	Rabattement s (cm)	Distance au puits r (cm)
1	203	300
2	108	500
3	42	800

On calcule aussi les quantités suivantes:

$\sum s_i = 362$	$\sum r_i = 1600$	$\sum \ln(r_i) = 18.6$	$\sum (s_i - \bar{s})(r_i - \bar{r}) = -39567$
$\sum s_i^2 = 56342$	$\sum r_i^2 = 980000$	$\sum (\ln(r_i))^2 = 115.8$	$\sum (s_i - \bar{s})(\ln(r_i) - \overline{\ln(r)}) = -78$
$\sum (s_i - \bar{s})^2 = 12661$	$\sum (r_i - \bar{r})^2 = 126667$	$\sum (\ln(r_i) - \overline{\ln(r)})^2 = 0.48$	SCE=3.19

a) Calculez le coefficient b_1 de la régression, et déduisez la transmissivité de l'aquifère exprimée en cm^2/s
b) Quel est l'intervalle de confiance de niveau 95% pour la transmissivité de l'aquifère?
c) On décide d'implanter 3 autres piézomètres aux mêmes distances mais dans une direction orthogonale à la direction définie par les 3 premiers. Indiquez comment on devrait utiliser la régression pour pouvoir tester si la transmissivité est la même selon les deux directions.

7- Le graphe suivant montre les résidus d'une régression en fonction de diverses variables. Indiquez dans chaque cas si le résultat est suspect ou non. Si vous trouvez le résultat suspect, indiquez ce que vous pourriez faire pour tenter d'améliorer le modèle.



8- On désire estimer la moyenne d'un relevé gravimétrique par un modèle de type polynômial afin de filtrer la dérive pour faciliter l'analyse par approche fréquentielle. On envisage 3 modèles possibles pour cette dérive:

(1) Gravi=b0+e
(2) Gravi=b0+b1*x+b2*y+e
(3) Gravi=b0+b1*x+b2*y+b3*x^2+b4*x*y+b5*y^2+e

On a 30 observations.

Les sommes des carrés des erreurs des différents modèles valent:

Modèle	SCE ($mgal^2$)
(1)	9000
(2)	3000
(3)	2800

a) Quel modèle devrait-on retenir pour estimer la moyenne spatiale de l'anomalie gravimétrique? Faites les tests requis au niveau 5%.
b) Donnez les unités des coefficients de la régression si x et y sont en m et Gravi en mgal.

9- On désire modéliser l'écoulement régional d'un aquifère de surface situé sur la rive sud de Montréal. Pour ce faire, on effectue le relevé de 60 piézomètres en un court laps de temps au mois d'avril (moment où la nappe est approximativement à son niveau maximal). On note les coordonnées *x, y* donnant la localisation en plan des piézomètres, *z* l'élévation du sol au piézomètre et *h* la charge hydraulique. On tente d'ajuster par régression divers modèles et l'on obtient les résultats suivants :

Modèle	Équation	SCE (m^2)
A	$h(x,y)=b_0+e$	70
B	$h(x,y)=b_0+b_1x+b_2y+e$	20
C	$h(x,y)=b_0+b_1x+b_2y+b_3z+ e$	10
D	$h(x,y)=b_0+b_1z+ e$	13

a) Quel modèle devrait-on retenir? Indiquez vos calculs et faites les tests nécessaires.

b) Que vaut le R^2 de ce modèle ?

On refait le même relevé, mais cette fois au mois de septembre. Pour la suite, on suppose que le modèle C est le modèle retenu.

c) Bien que le niveau moyen de la nappe en septembre soit certes inférieur à celui du mois d'avril, comment pourriez-vous vérifier néanmoins que la direction régionale d'écoulement demeure la même (d'un point de vue statistique) pour ces deux périodes de l'année ? Indiquez les modèles utilisés et les degrés de liberté associés au test. Aide : partant du modèle C construisez un modèle imposant les mêmes coefficients de régression pour x et y aux deux jeux de données (avril et septembre). Comparez ce modèle à un modèle permettant des coefficients différents pour x et y selon le relevé.

10- Lorsque l'on estime les réserves d'une mine, il est important de tenir compte des variations de densité de la roche. Souvent ces variations de densité sont liées assez directement à la teneur du minerai, surtout lorsque le minerai est riche et constitué de sulfures plus denses que la roche mère. Conscient de ce fait, un ingénieur géologue d'une mine de Cu-Zn (Cu dans la chalcopyrite, Zn dans la sphalérite) a mesuré avec précision la densité de 100 échantillons représentatifs de la mine. Il a de plus effectué l'analyse géochimique de ces mêmes échantillons. Le tableau suivant présente une partie des résultats qu'il a obtenus (d_i : densité de l'observation i (sans unité); $CuZn_i$: Cu+Zn pour l'observation i, en %) :

Statistique	Valeur
Moy(d)	3.1
Var(d)	17
Moy(CuZn)	4.2 %
Var(CuZn)	280 $\%^2$
Cov(d,$CuZn_i$)	28 %
SCT_m	1700

a) Utilisant ces résultats, et supposant que la densité du minerai est liée linéairement aux teneurs de Cu+Zn, donnez l'équation de prédiction de la densité du minerai.

b) Quelles sont les unités associées aux coefficients b_0, et b_1 de la régression ?

c) Que vaut le coefficient de corrélation simple ?

d) Que vaut le coefficient de détermination R^2 ?

e) Que vaut CME ?

f) Un minerai montre une teneur de 4% Cu et 5% Zn. Quelle est la valeur prédite de la densité?

g) En additionnant les teneurs en Cu et Zn pour effectuer sa régression, quelle hypothèse a été faite implicitement par l'ingénieur?

11-*Parmi les 4 modèles suivants, un seul peut, après transformation, être traité par régression linéaire. Indiquez quel est ce modèle et donnez son expression après transformation.*

i. $Y = b_0 + b_1X_1 + b_2X_2^{b_3} + e$

ii. $Y = b_0X_1^{b_1}e$

iii. $Y = b_0X_1^{b_1} + e$

iv. $Y = b_0 + b_1\text{Cos}(b_2X_1)+e$

12- *Dans une régression linéaire multiple, à quelle somme de carré le produit Y'Xb est-il égal ? Démontrez.*

13- Dans Excel, la fonction « Growth » (ou « Croissance » en français) permet d'estimer le modèle :

$$\hat{Y} = b_0 b_1^x$$

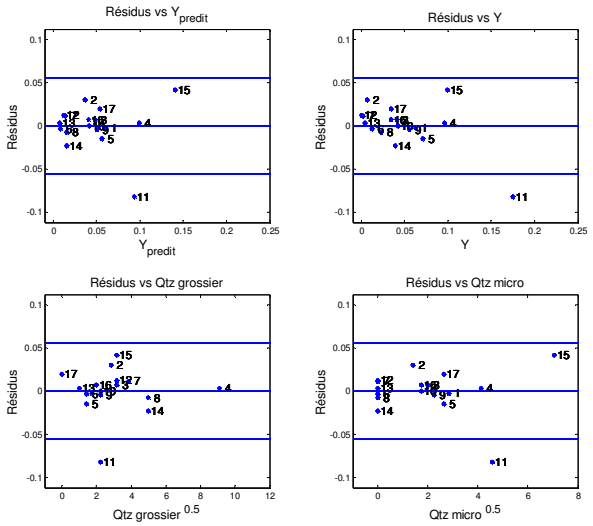
a) Indiquez comment vous pourriez « linéariser» ce modèle et ainsi obtenir des estimés pour b_0 et b_1 avec un programme de régression linéaire. Indiquez clairement le vecteur « y » et la matrice « x » qui seront soumis au programme de régression, les coefficients obtenus par la régression et le lien avec les coefficients recherchés.

b) Les prédictions $\hat{Y}$ *obtenues avec ce modèle minimiseront-elles la somme des carrés des erreurs (si l'on définit « e » comme* $e = Y - \hat{Y}$ *) ? Justifiez.*

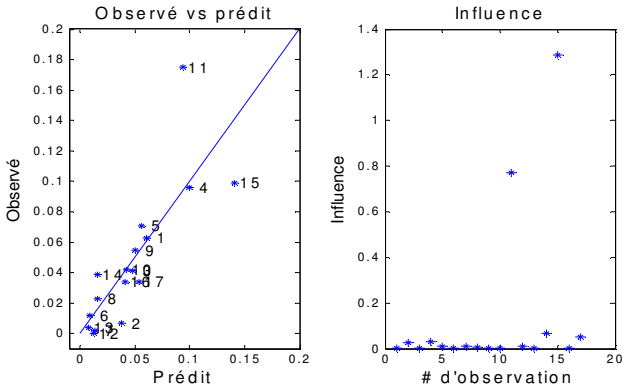
14- On a déterminé la porosité d'échantillons de résidus miniers provenant de 2 sites différents.

Expliquez comment on peut utiliser un programme de régression pour tester l'égalité des moyennes des porosités des 2 sites. (Aide : on doit utiliser le test d'ajout)

15- On a mesuré la déformation de 17 éprouvettes de béton après une cure humide d'un an. On a aussi déterminé le % de quartz micro-cristallin (chert et calcédoine) et le % de grains de quartz plus grossiers pour chaque éprouvette. Ces déterminations ont été réalisées par comptage microscopique. L'objectif de la régression est de prédire la déformation après un an en utilisant les % de quartz comme variables prédictives. Les graphiques suivants montrent les résidus obtenus en fonction de certaines variables. Les limites correspondant à + ou – 2 *$\text{CME}^{0.5}$ sont aussi illustrées :

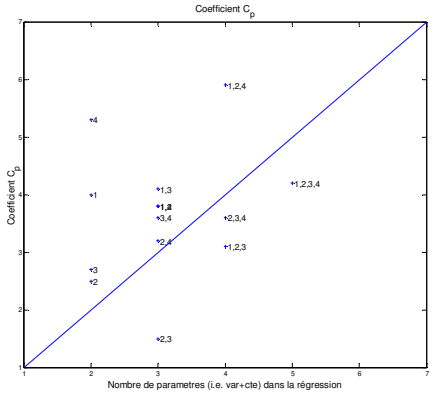


Le graphique suivant montre le graphe des valeurs observées vs prédites et celui de l'influence des observations :



Les résultats présentés vous semblent-ils conformes à ce qui est attendu ? Justifiez votre réponse.

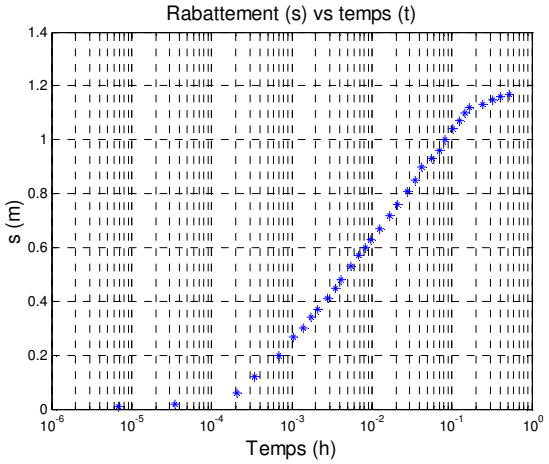
16- Soit le diagramme suivant illustrant le coefficient C_p obtenu pour différents sous-ensembles de variables.



Quel ensemble de variables ce graphique vous suggère-il retenir? Justifiez.

17- Dans un aquifère à nappe confinée, homogène, de grande extension latérale et d'épaisseur constante, l'on dispose d'un piézomètre ayant une crépine sur toute l'épaisseur de l'aquifère. On mesure le rabattement (s) en fonction du temps en vu d'estimer la transmissivité (T) et le coefficient d'emmagasinement (S) à l'aide de la méthode de Cooper-Jacob qui consiste à

ajuster une droite sur la partie linéaire du graphe s vs ln(t). Le graphique suivant montre le rabattement en fonction du temps.



Suggérez un outil vu au cours qui pourrait permettre d'identifier et d'éliminer de façon itérative les points qui s'éloignent trop de la partie droite de la courbe.

18- On veut utiliser une sonde mesurant la conductivité électrique en continu sur un convoyeur pour identifier les fragments minéralisés et les fragments non-minéralisés en vu de les séparer. La teneur économique pour le gisement est 0.5% Ni équivalent (le seuil définissant « minéralisé »). On définit le Ni équivalent comme Ni+0.8*Cu. Vous avez mesuré la teneur chimique en Ni et en Cu sur 100 fragments et avez mesuré sur ces mêmes fragments la conductivité électrique. Vous notez une relation linéaire assez forte entre le log(conductivité) et le Ni équivalent.

Quel modèle de régression utiliserez-vous pour rencontrer les objectifs de l'étude?

19- Vous travaillez comme stagiaire chez Hydro-Québec (HQ) sur un chantier où vous êtes chargés d'effectuer le contrôle de vibrations dues à des tirs de sautage effectués par un entrepreneur indépendant. Votre séismographe enregistre la vitesse de déplacement des particules (v). Vous connaissez la distance par rapport au tir (d) et la charge exacte utilisée pour le tir (w). HQ utilise un modèle de la forme suivante pour faire les prédictions des vitesses :

$$\ln(v) = b_0 + b_1 \ln\left(\frac{d}{w^{1/2}}\right) + e \qquad \text{(premier modèle)}$$

a) Quel est le signe attendu du coefficient « b_1 » ?

b) Décrivez le vecteur Y et les colonnes de la matrice X correspondant à ce problème de régression.

Après un certain nombre de relevés, vous vous demandez si le modèle utilisé par HQ est le meilleur modèle possible. Vous considérez le modèle alternatif suivant :

$$\ln(v) = b_0 + b_1 \ln(d) + b_2 \ln(w) + e \qquad \text{(second modèle)}$$

c) Décrivez le vecteur Y et les colonnes de la matrice X correspondant à ce problème de régression.

d) Avec les mêmes 30 observations, vous obtenez R^2 =0.8 avec le premier modèle (SCE=2) et R^2=0.9 avec le second modèle. Le second modèle est-il significativement meilleur que le premier ? Posez l'hypothèse H_0 à vérifier et faites le test requis.

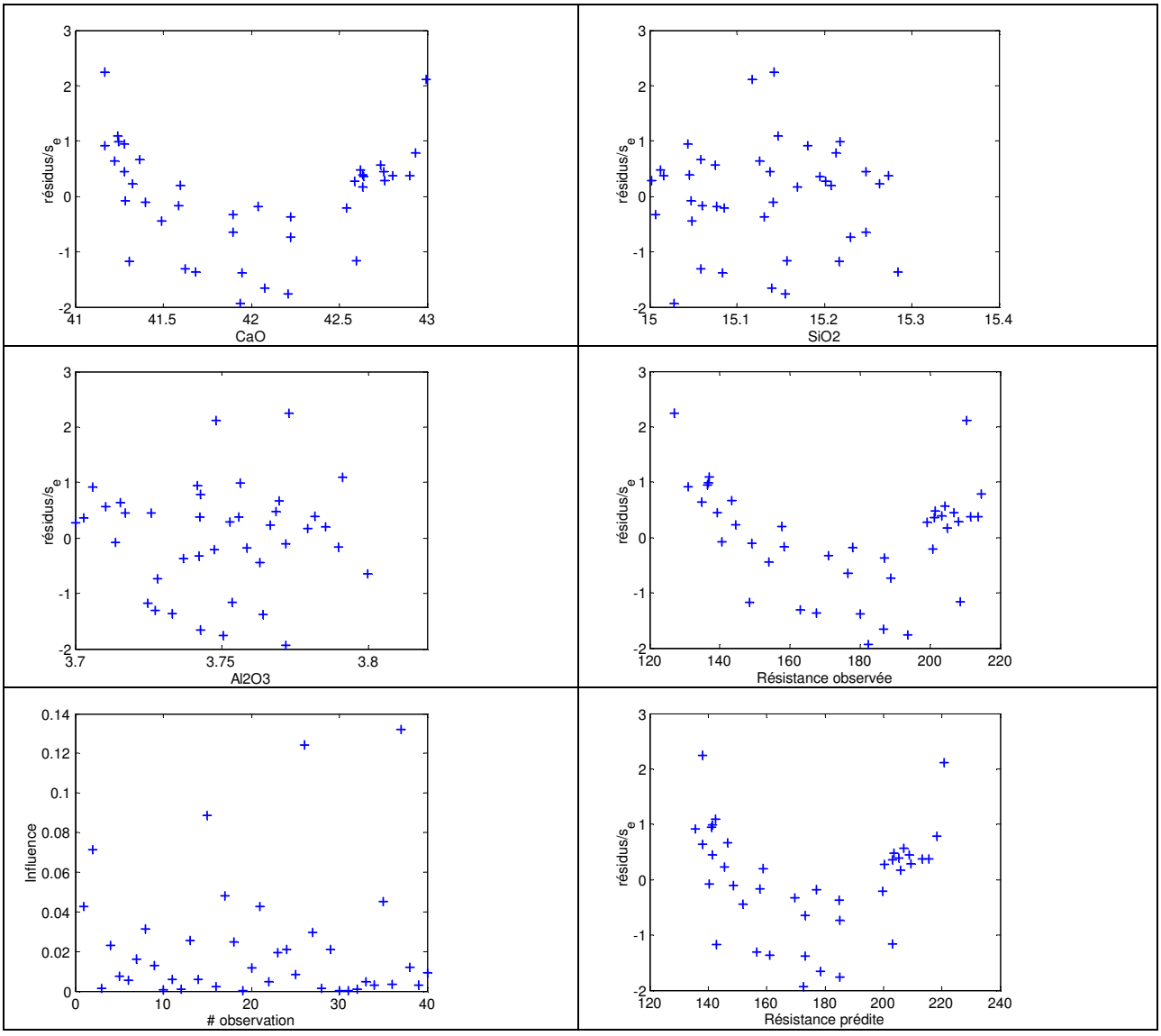
e) L'entrepreneur effectue ses propres mesures de vibrations. Il adopte un modèle identique à votre second modèle (i.e. il adopte un modèle de la forme $\ln(v) = b_0 + b_1 \ln(d) + b_2 \ln(w) + e$ *) mais trouve ses propres coefficients « b ».*

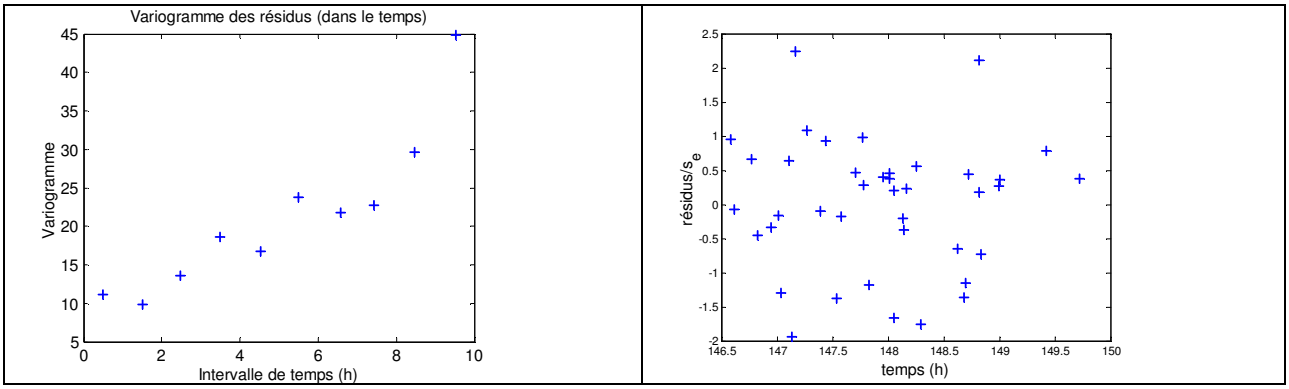
Expliquez comment vous pourriez tester si ces coefficients sont significativement différents des vôtres. Donnez tous les détails requis (vecteur Y, matrice X, sommes des carrés à utiliser, degrés de liberté du test.)

Dans les opérations quotidiennes, vous ne connaissez pas la charge « w » réellement utilisée par l'entrepreneur. Le rôle du contrôle des vibrations est justement d'assurer que l'entrepreneur n'utilisera pas des charges excessives (afin d'accélérer les travaux), à l'insu de HQ, risquant ainsi d'endommager le roc (une règle interne de HQ indique que la vitesse des particules à 20m ne doit pas excéder 5cm/s. Tout tir excédant cette règle doit être rapporté à HQ). Vous relevez la vitesse des particules en positionnant le géophone à une distance variant de 30m à 100m du lieu de tir selon la disponibilité des sites et la sécurité de l'opération. Ayant la vitesse des particules et le modèle précédent (second modèle), vous déterminez la charge « ln(w) » utilisée lors du tir. Ayant cette charge et le modèle, vous calculez la vitesse des particules (ln(v)) que vous auriez enregistré si votre géophone avait été positionné à 20m du tir.

f) Identifiez un problème qui se pose avec cette approche.

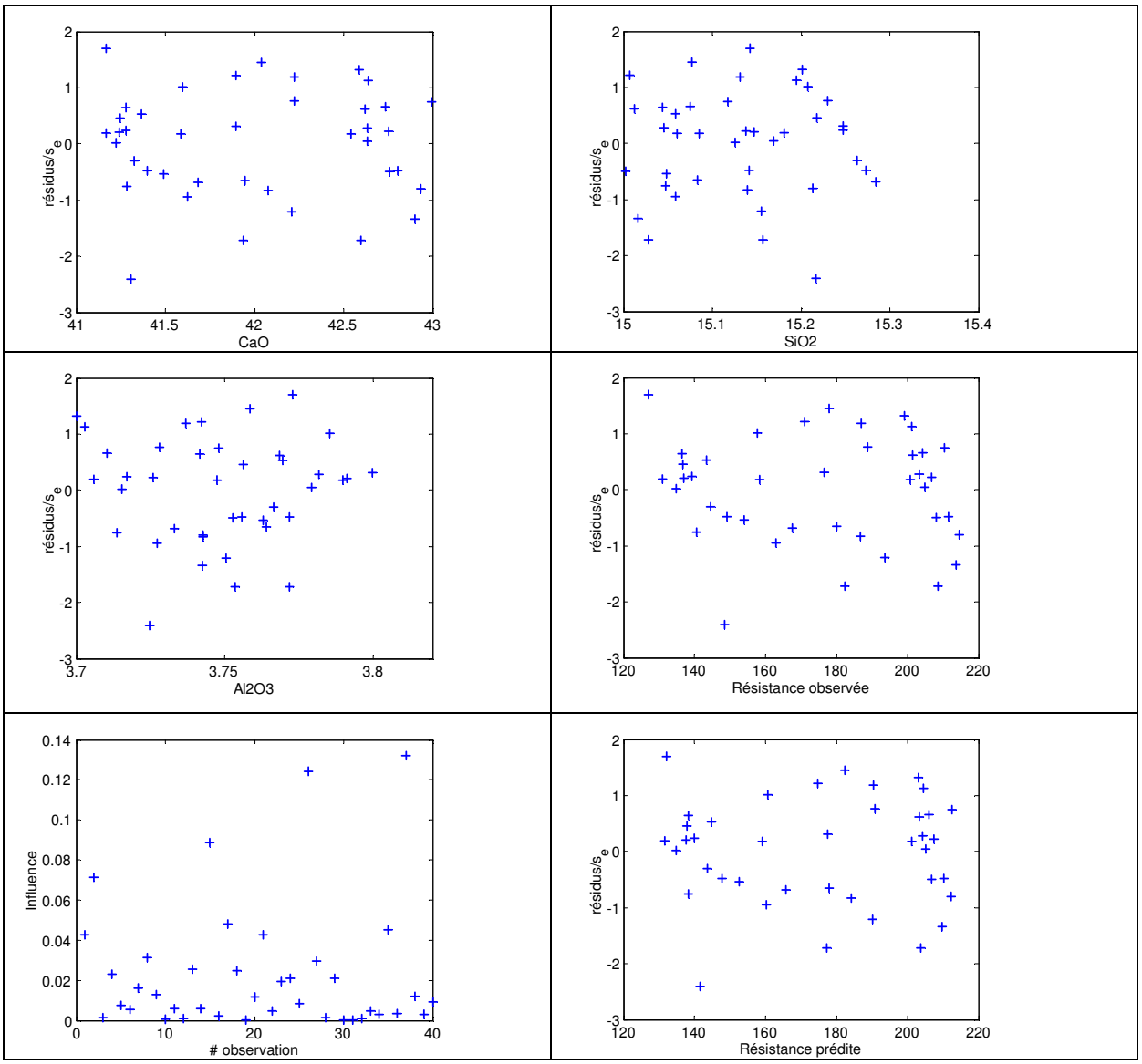
20- On veut prédire la résistance d'un ciment (kPa) à partir de la composition chimique du calcaire prélevé à la carrière. On établit un modèle de prédiction puis on calcule les résidus de ce modèle. Voici 8 graphes choisis des résidus :

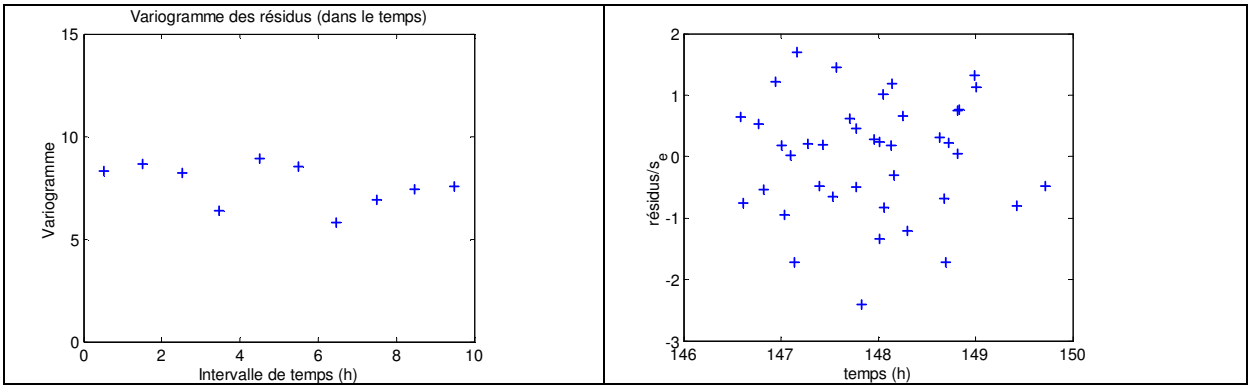




a) Identifiez tout comportement des résidus qui vous semble suspect en indiquant le problème suspecté.

On effectue une certaine correction pour améliorer le modèle et voici maintenant les graphes des résidus obtenus :





b) La correction apportée a-t-elle réussi à corriger le(s) problèmes(s) identifié(s) précédemment ? Discutez.

21- Indiquez pour les modèles suivants s'ils sont i. linéaire, ii. non-linéaire mais pouvant être rendu linéaire par transformation, ou iii. non-linéaire.

a) $Y = X_1^{\beta_1} X_2^{\beta_2} \varepsilon$

b) $Y = \beta_0 + \beta_1 X_1 X_2^{3.1} + \beta_2 X_2 + \varepsilon$

c) $Y = \beta_0 \beta_1^{X_1} \varepsilon$

d) $Y = \beta_1 cos(\beta_2 X_1) + \varepsilon$

22- Lors de tirs explosifs sur un chantier de construction, un séismographe enregistre la vitesse de déplacement vertical des particules (v en m/s). Vous connaissez la distance entre le séismographe et le tir (d en m) et la charge explosive utilisée (w en kg). Vous effectuez 20 mesures et construisez un modèle de la forme suivante pour faire les prédictions de ln(v) :

$$ln(v) = \beta_0 + \beta_1 ln\left(\frac{w^{1/2}}{d}\right) + e \quad \text{(modèle A)}$$

a) *On obtient un R^2 de 0.9, SCE vaut 110. Le coefficient β_1 est-il significatif ? Faites le test requis pour vous en assurer.*

Vous effectuez une régression cette fois avec le modèle suivant :

$$ln(v) = \beta_0 + \beta_1 ln\left(\frac{w^{1/2}}{d}\right) + \beta_2 ln(d) + e \quad \text{(modèle B)}$$

b) Le R^2 passe à 0.93. Faites le test permettant de juger si le modèle B est significativement meilleur que le modèle A.

Le client du projet de construction exige de l'exécutant des travaux que les explosions génèrent une vitesse de déplacement vertical inférieure à 100m/s pour un point situé à 20m de distance du point de sautage (ceci pour éviter des endommagements aux structures existantes et au roc de fondation).

c) Expliquez comment vous pourriez utiliser le modèle B pour indiquer à l'exécutant la quantité maximale de charge explosive (w_{max}) pouvant être utilisée pour respecter la norme fixée par le client dans la grande majorité des cas.

(Aide : vous devez accepter un certain niveau de risque de dépasser le seuil fixé; ce risque porte sur la vitesse de chaque tir individuellement et non sur la valeur moyenne de la vitesse pour une charge donnée).

Utilisant le même jeu de données, on calcule cette fois les paramètres du modèle de régression :

$$ln(w) = \beta_0 + \beta_1 ln(d) + \beta_2 ln(v) + e \qquad \text{(modèle C)}$$

d) Votre employeur vous demande en cours de construction de mesurer les vitesses observées, la distance au tir et d'en déduire la charge explosive utilisée pour s'assurer que le constructeur respecte la charge maximale déterminée en c) (Note : l'exécutant peut être tenté d'accroître la charge pour accélérer les travaux). *Quel modèle utiliserez-vous pour faire ce travail, le modèle B ou le modèle C? Pourquoi?*

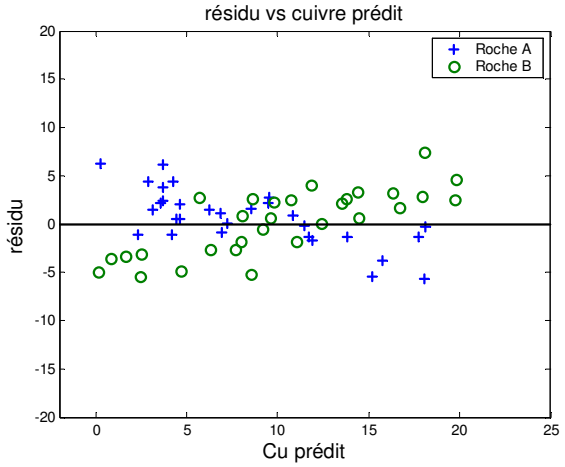
23- On veut prédire la teneur en Cu d'une roche à partir des mesures de la conductivité électrique et de la densité obtenues par sondes géophysiques. Le modèle actuel s'écrit :

$$Cu = \beta_0 + \beta_1 \rho + \beta_2 d + \varepsilon$$

où Cu est le % de Cu à l'analyse géochimique, ρ est la résistivité apparente (en ohm-m) et *d* la densité (sans unité). Le Cu se trouve dans des sulfures qui ont la caractéristique d'être plus denses et plus conducteurs que la roche encaissante.

a) Quel est le signe attendu des coefficients b_1 et b_2 ? Quelles sont les unités des coefficients b_1 et b_2 ?

Deux types différents de roche (A et B) sont utilisés dans cette expérience. Le graphe des résidus en fonction des valeurs prédites est construit en utilisant un marqueur différent pour chaque type de roche. On obtient :



b) Indiquez ce qui cloche au niveau des résidus. Suggérez un modèle alternatif au modèle actuel qui pourrait corriger le problème observé (utilisant la même information que disponible présentement et traitant toutes les données en un seul bloc) .

Dans le but de pouvoir distinguer dans les sulfures la portion pyrite de celle chalcopyrite, on a effectué l'analyse géochimique du Fe en plus du Cu. On a obtenu la matrice de corrélation simple suivante :

	Cu	Fe	ρ	d
Cu	1.000	0.285	-0.873	0.734
Fe	0.285	1.000	-0.684	0.786
ρ	-0.873	-0.684	1.000	-0.906
d	0.734	0.786	-0.906	1.000

c) Quel serait le R^2 d'un modèle prédisant le Cu uniquement avec la résistivité ?

24- On veut effectuer la correction géométrique d'une image satellite déformée. Soit (u,v) les coordonnées sur l'image déformée et (s,t) les coordonnées sur l'image de référence. On réussit à identifier 10 points de contrôle sur l'image de référence et sur l'image déformée. Un modèle polynomial en (s,t) est utilisé pour corriger géométriquement l'image. Pour la suite, on ne s'intéresse qu'à la régression portant sur « u ».

a) La régression $u^*=b_0+b_1s+b_2t$ fournit un SCE de 10. Lorsqu'on considère le modèle quadratique, $u^*=b_0+b_1s+b_2t+b_3s^2+b_4s^*t+b_5t^2$, on obtient SCE=8. L'inclusion de la composante quadratique améliore-t-elle significativement le modèle ? Faites le test pour vous en assurer.

b) Quel serait le degré maximum du polynôme que l'on pourrait considérer utiliser dans ce problème en supposant que tous les coefficients du polynôme sont présents dans la régression ? Pourquoi serait-il risqué de retenir un polynôme de degré aussi élevé ?

c) On calcule avec le modèle linéaire de la question a) $\sum_{i=1}^{10} u_i u_i^* = 100$ où u_i^* est la valeur prédite pour le $i^{ième}$ point de contrôle. Que vaut SCT ?

25-Vous inspirant de l'exemple vu en classe portant sur les intervalles de confiance simultanés (ellipse ou ellipsoïde de confiance sur l'ensemble des coefficients), et les intervalles de chaque coefficient considéré individuellement,

a) est-il possible que tous les coefficients de la régression soient individuellement significatifs mais que la régression dans son ensemble ne soit pas significative ?

b) est-il possible que tous les coefficients, individuellement soient non-significatifs mais que la régression dans son ensemble soit significative ?

26- Vous réalisez une régression avec des données provenant de 3 laboratoires différents (modèle : Y vs une seule variable X_1). Après examen des résidus, vous constatez qu'il y a un problème et qu'il existe un effet lié aux laboratoires.

Construisez un modèle unique (i.e. toutes les données sont traitées simultanément en une seule régression) permettant de définir trois droites de régression différentes (ordonnée à l'origine et pente). Indiquez clairement la signification des variables et dites comment obtenir les pentes et les ordonnées à l'origine des 3 droites à partir des différents coefficients de la régression.

27-*.Dans un modèle de régression avec constante, quel est le coefficient D_i d'une observation dont la valeur pour chaque variable est égale à la moyenne de cette variable sur l'ensemble des observations (.i.e $x_{i,1} = \overline{x}_1 \quad x_{i,2} = \overline{x}_2 \quad ....x_{i,p} = \overline{x}_p$) ?*

Justifiez.

28- On effectue des mesures de perméabilité (k) à n=57 profondeurs différentes le long d'un puits entièrement crépiné localisé dans des dépôts meubles. On effectue aussi des relevés aux mêmes points le long du puits à l'aide de 3 sondes géophysiques (gamma, neutron et électromagnétique). La sonde gamma mesure la quantité d'argile dans le dépôt (gamma augmente avec la quantité d'argile), la sonde neutron donne la porosité et la sonde électromagnétique donne la résistivité.

On veut expliquer le $\log_{10}(k)$ à l'aide des variables obtenues des sondes géophysiques et la profondeur du levé. On effectue la sélection avant et on obtient :

Variable	R^2	SCE	F (ajout)
gamma (cps)	0.268	3.749	20.14
porosité (fraction)	0.353	3.312	7.13
résistivité (ohm-m)	0.394	3.106	3.51
profondeur (pi)	0.394	3.105	0.02

Selon ces résultats quelles variables devrait-on retenir dans le modèle de régression? Faites les tests requis.

29- Dans un gisement de cuivre, la quantité de sulfures (%) dans une roche contrôle la résistivité (ohm-m) de celle-ci. La mine veut prédire la teneur en sulfures de son minerai à partir d'une mesure de la résistivité que l'on ferait en continu à l'entrée du concentrateur. On considère les deux modèles suivants :

Modèle A : sulfures = $b_0 + b_1$ résistivité +e

Modèle B : résistivité = $c_0 + c_1$ sulfures +e

a) Quelles sont les unités des coefficients b_1 et c_1 ?

b) Le coefficient b_1 du modèle A vaut 0.5. La variance des concentrations en sulfure vaut $10\%^2$. La variance de la résistivité vaut 7.6 $(ohm\text{-}m)^2$. Quelle est la valeur du coefficient c_1?

c) Quel modèle devrait-on utiliser compte tenu de l'objectif poursuivi ?Justifiez.

Corrigé

Questions sur les tests

1- On calcule Q=$(10-9)^2/9+(55-50,4)^2/50,4+...=1,32$. On compare à une $\chi^2_{4-0-1,0.05} = 7,815$. On ne peut pas rejeter H_0, donc la distribution proposée semble plausible.

2- a) H_0 p=0.6 vs H_1 p>0.6.

On rejette H_0 si $\frac{(\hat{p}-0.6)\sqrt{1000}}{(0.4*0.6)^{1/2}} > 1.645$. On calcule $\frac{(\hat{p}-0.6)\sqrt{1000}}{(0.4*0.6)^{1/2}} = 2,9$, donc on rejette H_0.

b) D'une table N(0,1) on tire que le niveau α correspondant à 2.9 est 0.0019.

3-

a) Sous H_0, $\mu = 3.5$ on calcule $(4.2-3.5)/(1.96/9)^{0.5}=1.5 <1.645=Z_{.05}$ Donc on ne peut pas rejeter H_0. On ne peut pas invalider l'affirmation du fabricant.

b) Sous H_1, $\mu = 3.8$, On cherche $P(\bar{X} > 1.645(1.96/9)^{0.5} + 3.5) = P(\frac{\bar{X}-3.8}{(1.96/9)^{0.5}} > 1.645 + \frac{3.5-3.8}{(1.96/9)^{0.5}})$= P(Z>1.0)=0.1587

c) $n \geq (z_{0,05}+z_{0,10})^2 1,96/(3.8-3.5)^2 = 186,6 => 187$ cigarettes

---

4- a) H_0 $\sigma_1^2 = \sigma_2^2$ vs H_1 $\sigma_1^2 \neq \sigma_2^2$. On calcule $\bar{x}_1$=305, $\bar{x}_2$=316, s_1^2=102, s_2^2=8

$s_1^2/s_2^2=102/8=12,75<F_{3,3,;0,025}=15,4$. On ne peut rejeter l'hypothèse que les bétons aient la même variance. (Note : ici on utilise un test bilatéral car on n'a pas d'a priori sur les relations entre les variances).

b) On fait le test en supposant $\sigma_1^2 = \sigma_2^2$. On test H_0 $\mu_1 = \mu_2$ vs H_1 $\mu_1 < \mu_2$

On calcule $s_p^2=(3s_1^2+3s_2^2)/(4+4-2)=55$

$\frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}} = -2,10$ à comparer à $t_{4+4-2;\,0,05}$= -1,942

Comme –2,10<-1,942, on rejette H_0, donc le 2e béton est significativement plus résistant que le premier.

5- a) P(Rejeter H_0 |H_0 est vraie) = $P(X_1>0,9 \mid X\sim U(0,1))$= 0,1

b) P(Rejeter H_0 | H_1 vraie) = $P(X_1>0,9 \mid X\sim U(0,1.5)=0,6/1,5=0,4$

6- a) On rejette H_0 si $\bar{X}$ < -1,645*5/3+50=47,25

P(rejeter(H_0|X~N(45,25)=P(X<47,25)=P(Z<(47,25-45)*3/5)=P(Z<1.35)=0.91

b) $n \geq (z_{0,05}+z_{0,10})^2 25/(45-50)^2=(1,645+1,28)^2=8,6 => 9$

7- Sous hypothèse d'indépendance, les fréquences théoriques sont :

	Revenu			
Succès	Faible	Moyen	Élevé	Total
Faible	8	24	8	40
Moyen	12	36	12	60
Élevé	10	30	10	50
Total	30	90	30	150

On calcule Q=2,8> $\chi^2_{4,0,05}$ =9,49. On ne peut pas rejeter H_0.

8- On calcule Q=2,8> $\chi^2_{4-0-1;0,05}$ =7,81. On ne peut pas rejeter H_0.

9- a) n=$(z_{0,025}+z_{0,10})^2 \frac{\sigma^2}{(u_0-u_1)^2}$ =$(1,96+1,28)^2 4/1^2$=42,04 =>43

b) $z_{calculé}$=(65-64,3)*$43^{1/2}$/2=2,29 > $z_{table,0,025}$=1,96, donc on rejette H_0.
c) le seuil critique de rejet est 64,3+1,96*2/sqrt(43)=64,9. On cherche P($\overline{X} < 64.9 \mid \mu = 65$) =P(Z<(64.9-65)*sqrt(43)/2)= P(Z<-0,3279)=0,3715
d)On veut P($\overline{X} > 65 \mid \mu = 64,3$) =0,05. P(Z>(65-64.3)*sqrt(n)/2)=0,05 => >(65-64.3)*sqrt(n)/2=1,96 =>n=31,36=>32

10- On calcule s_1^2=5,0, s_2^2=3,9048. Le ratio s_1^2/s_2^2=1,28<$F_{6,6;0,05}$=4,28. On accepte H_0.

11- On a : 22,72/14,91=1,52<$F_{9,9;0,05}$=3,18. On accepte H_0.
On calcule s_p^2=(9*22,72+9*14,91)/18=(22,72+14,91)/2=18,82 et s_p=4,34.
H_0 : $\mu_2 = u_1$ vs H_1 $\mu_2 \neq u_1$
$z_{calculé}$= $\frac{55,72-50,29}{4,34\left(\frac{1}{10}+\frac{1}{10}\right)^{0.5}} = 2,80$ $t_{18;0,05}$=2,10
Comme 2,8>2,1, on rejette H_0 et l'on conclut que l'entreprise 2 fournit des tiges plus résistantes que l'entreprise 1.

Questions supplémentaires sur les régressions

1- On calcule :

a) x est la mesure au Courrier, y est l'analyse chimique
moy(x)=241/5=48.2
moy(y)=260/5=52
Sxy=(12617- 5*48.2*52)=85
Sx^2=11700-5*48.2^2=83.8
b1=Sxy/Sx^2=85/83.8=1.0143
b0=moy(y)-b1*moy(x)=52-1.0143*48.2=3.110

La valeur corrigée sera donc: 3.11+1.0143*50=53.8%

b) SCR=Yp'*Yp=Y'*Yp=Y'Xb=260*3.1107+12617*1.0143=13606.2
SCM=n*$moy(y)^2$=5*52^2=13520
SCRm=13606.2-13520=86.2
SCTm=13618-13520=98
SCE=98-86.2=11.8
R2=86.2/98=0.88

Autre approche :
r=Sxy/(Sx*Sy)=85/$(98*83.8)^{0.5}$=0.93796
R^2=0.88
SCRm=0.88*SCTm=0.88*98=86

Le reste est identique.

Test (86/1)/(12/3)=86/4=21.5 >F(1,3 ;.05)=10.1 La régression est significative.

2- Ici n=22,p=3. CME=SCE(avec toutes les données)/(n-p-1)=249/18=13.83. L'influence est donc 4/(4*13.83)=0.07
L'influence de cette observation n'est pas anormalement forte, il n'y a pas lieu de s'inquiéter.

3- Il suffit de faire un test d'ajout. Le modèle réduit est le modèle avec seulement la constante b0. On définit une variable indicatrice prenant la valeur 1 si l'observation vient du site 1 et 0 sinon (ou l'inverse). On effectue la régression avec cette variable indicatrice. Si l'ajout est significatif c'est que les 2 sites n'ont pas la même moyenne.

4- On effectue la régression modèle complet modèle réduit. La meilleure sonde au départ est la 1ère puisque le R2 est 0.8.
SCTm=112*38=4256
SCRm(r)=0.8*4256=3404.8
SCRm(c)=0.85*4256=3617.6

Test d'ajout : [(3617.6-3404)/2]/[(4256-3617.6)/33]=5.52
Le F (2,33 ;5%) vaut 3.32, on considère que l'ajout est significatif.

5- On peut écrire $Y-\overline{Y}=b_1(X-\overline{X})+e$ qui est un modèle sans constante.

6- a) b1=-78/0.48=-162.5 cm
$T=\frac{-Q}{2\pi b_1}$=-100/(2*pi*-162.5)=0.098 cm^2/s
b) Var(b1)=Var(e)/0.48 (voir notes p.13)

Var(e)=SCE/1=3.19
Var(b1)=6.64

Intervalle sur b1 : +- t(1;0.95) se= +- 12.7*2.58= =- 32.8

b1 compris entre [-195.3,-129.7] T entre [.081 , .122]

c) On ferait la même régression qu'en b) avec les 6 observations et on noterait le SCE correspondant. On ferait ensuite la régression en codant une des 2 directions et en utilisant le modèle:
s=b0+b1*ln(r)+b2*I*ln(r) avec I=1 si une des direction et 0 pour l'autre. On teste ensuite le caractère significatif de b1 ou, ce qui est la même chose, on fait le test d'ajout. Si on rejette le caractère significatif alors on peut conclure que les 2 directions montrent la même transmissivité, sinon il existe une différence significative.

7- a) La tendance linéaire provient d'un mauvais choix en abscisse. Il aurait fallu mettre la valeur prédite (comme en b). Il est normal d'observer un lien linéaire entre résidus et valeurs observées.
b) Normal
c) Il y a une tendance linéare dans le temps. Il faudrait inclure cette variable dans le modèle.
d) Il y a une tendance quadratique en fonction de la variable x. Il faudrait inclure, en plus de x une variable x^2 ou peut-être $x^{0.5}$ ou du moins une transformation qui permette de ramener le graphe à une bande homogène autour du niveau 0.

8- (2) vs (1) : $\frac{(9000-3000)/2}{3000/(30-3)}=27$ vs F(2,27;0.05)=3.35 Fortement significatif

(3) vs (2) $\frac{(3000-2800)/3}{2800/(30-6)}=0.57$ vs F(3,24,0.05)=3.01 Non-significatif. On retient la dérive linéaire.

b) Coefficient pour b0: mgal; b1: mgal/m; pour b2: mgal/m

9- a) Test d'ajout ; on peut tester D par rapport à A puis C par rapport à D.
D vs A : F calculé : (57/1)/(13/58)=254 Fortement significatif
C vs D : F calculé : (3/2)/(10/56)=8.4 > 3.16=F(2,56 ;0.95) Significatif on conserve C

b) SCTm=70, SCE=10, SCRm=70-10=60, R^2=60/70=0.86.

c) Il suffit de mettre les 2 jeux de données ensemble et de tester le modèle

F : $h(x,y)=b_0+b_1x+b_2y+b_3z+b_4I+b_5Iz+e$
vs
G: $h(x,y)=b_0+b_1x+b_2y+b_3z+b_4I+ b_5Iz +b_6Ix+b_7 Iy+e$

En testant l'ajout de G vs F, on teste l'égalité à 0 de b_5 et b_6. Si l'on accepte l'hypothèse alors les directions régionales d'écoulement sont les mêmes, sinon il y a changement significatif. Le test comporte 2 degrés de liberté au num,érateur et 60-8=52 au dénominateur.

10- a) $b_1=s_{xy}/s_x^2$= 28/280=0.1 $(\%)^{-1}$
b_0=3.1-0.1*4.2=2.68

b) b0 est sans unité et b1 est en $\%^{-1}$

c) $r=28/(17*280)^{0.5}$=0.4058

d) R^2= 0.4058^2=0.1647

e) SCRm=R^2*SCTm=1700*0.1647=280
SCE=1700-280=1420
CME=1420/(100-2)=14.5

f) valeur prédite: 2.68+9*0.1=3.58%

g) que les coefficients de la régression pour ces 2 variables étaient égaux et donc que la densité de ces 2 minéraux (sphalérite et chalcopyrite étaient égales)

11- C'est le modèle ii. On prend le log : $\ln(Y)=\ln(b_0)+b_2 \ln(X_2) +\ln(e)$

12- $Y'Xb=Y'X(X'X)^{-1}X'Y=Y'MY=SCR$

13- a) En prenant les logs, on obtient : $ln(\hat{Y}) = ln(b_0) + ln(b_1) * x$. On a donc une régression linéaire avec constante de ln(Y) prédit par x. On obtient c_0 et c_1 de cette régression et $b_0=\exp(c_0)$ $b_1=\exp(c_1)$.

b) Comme la régression s'effectue dans un espace transformé, elle minimise la somme du carré des erreurs de la variable transformée ln(Y) et non du Y lui-même. Si l'on veut minimiser la somme du carré des erreurs, il faut effectuer une régression non-linéaire.

14- On forme un premier modèle réduit $Y=b_o+e$. On forme un second modèle complet : $Y=b_0+b_1*I +e$ où I est une variable indicatrice prenant la valeur 0 pour un site et 1 pour l'autre. Ce modèle permet d'avoir deux moyennes différentes comme estimation pour les deux sites alors que le modèle réduit ne permets qu'une seule moyenne. Il ne reste qu'à tester le caractère significatif de l'ajout. Si c'est significatif alors les moyennes diffèrent, significativement, sinon, les moyennes peuvent être considérées égales.

15- On note que l'observation 15 montre une trop grande influence. De plus l'observation 11 montre aussi une grande influence et son résidu est à l'extérieur de l'intervalle de confiance calculé avec $CME^{0.5}$. Cette observation est donc aussi très suspecte. Il faudrait possiblement refaire l'analyse sans ces 2 observations.

16- Tous les ensembles ayant les variables x_2 x_3 sont sous la diagonale. Parmi ceux-ci le couple (2,3) est celui le plus sous la

diagonale. Il s,agit probablement du meilleur sous-ensemble à retenir.

17- La notion d'influence d'une observation peut permettre de détecter les observations s'écartant de la droite. En enlevant les observations les plus influentes et en faisant le suivi du R^2 obtenu, on pourra déterminer un sous-ensemble de points situés sur une droite.

18- Le modèle aura la forme : Ni équivalent=b_0+b_1 log(conductivité électrique)+e

---

19- a) signe négatif car la vitesse décroît avec la distance et augmente avec la charge utilisée
b) Y : ln(v)
X : colonne de 1, colonne avec $d/w^{0.5}$.
c) Y : ln(v)
X : colonne de 1, colonne avec ln(d) et colonne avec ln(w)

d) On a R^2=0.8 =1-SCE/SCT_m=1-2/SCT_m. On trouve SCT_m=2/(1-0.8)=10.
SCR_m=SCT_m-SCE=10-2=8.

Avec le modèle complet, On a $SCR_{m,c}$=0.9*10=9. SCE_c=10-9=1
Le test est donc (2-1)/1 / 1/(30-3) = 27 > $F_{1,27;.05}$=4.21. L'ajout est très significatif.

e) il suffit de combiner les 50 observations et de considérer le second modèle comme modèle réduit et d'ajouter 3 variables permettant une droite de régression différente pour chaque cas : I, I*ln(d) I*ln(w). On teste l'ajout de ces 3 coefficients simultanément par rapport au modèle précédent. Le test comprend au numérateur la différence des SCE entre modèle réduit et complet avec 3 degrés de liberté. Au dénominateur, c'est SCE du modèle complet avec maintenant 30-6=24 degrés de liberté.

f) le problème est lié à la détermination de « w ». Les 2 modèles de régression sont construits pour prédire ln(v) et non ln(w). Il ne s'agit donc pas de l,estimation la meilleure que l'on puisse faire de ln(w) avec les données disponibles. Il faudrait faire une régression de ln(w) en fonction de ln(v) et ln(d) pour prédire de façon optimale ln(w).

20- a) i. On note une structure en arche sur le diagramme résidus vs CaO (inclure $CaO^{0.5}$ ou CaO^2)?
ii. on note une structure en arche sur le diagramme résidus vs valeurs prédites
iii. on note que le variogramme des résidus n'est pas un effet de pépite pur dans le temps. Il y a une composante temporelle manquante.

Note : le graphe résidus vs valeurs observées n'est pas comme attendu mais on ne doit pas l'utiliser pour établir le diagnostique.

b) tout est beau maintenant, le variogramme des résidus est un effet de pépite pur et les structures en arche sont disparues. Aucune observation n'a d'influence trop grande.

21- a) non-linéaire mais peut être linéarisé par transformation logarithmique (si X positif partout)
b) linéaire
c) non-linéaire mais peut être linéarisé par transformation logarithmique.
d) non-linéaire

22- a) R^2=0.9=1-SCE/SCT_m. donc SCT_m=SCE/(1-R^2)=110/0.1=1100. SCR_m=SCT_m-SCE=1100-110=990.
H_0 $\beta_1 = 0$ => test : (SCR_m/1) / (SCE/(20-2))=990/(110/18)= 162. La valeur $F_{table,\ 1,18,.05}$=4.41. Le coefficient est fortement significatif.
b) R^2=0.93=1-SCE/SCT_m; SCE=(1-R^2)SCT_m=0.07*1100=77.
Le test d'ajout est ((110-77)/1) / (77/(20-3))= 7.28 alors que $F_{tabl,1}$, 17,.05=4.45. Le modèle B améliore significativement la prédiction.

c) Ayant les coefficients de la régression, on peut fixer la distance à 20m, et tracer un graphe donnant simplement ln(v) en fonction de ln(w). On peut tracer la limite supérieure de l'intervalle de confiance de niveau $1-\alpha=95\%$ (ou 99% ou une autre valeur raisonnable) pour une valeur observée (p. 11)

$$Y_p \pm t_{n-(p+1),\alpha}\ s(1+\mathbf{x_i}(X'X)^{-1}\mathbf{x_i}')^{0.5}$$

On trace une ligne horizontale sur ce graphe à ln(100)=4.61. Le point d'intersection avec la limite supérieure indique la charge maximale que l'on peut utiliser tout en encourant un risque $\alpha/2$ de néanmoins dépasser la norme demandée.

d) Le modèle C. Clairement l'objectif ici est de prédire ln(w). C'est donc cette variable qui doit être le Y de la régression. On ne peut utiliser l'autre régression pour faire des prédictions. Comme en c), pour éviter de produire de fausses alarmes, on pourrait construire la borne supérieure de l'intervalle de confiance autour de chaque valeur prédite de ln(w) pour d=20. Si l'intervalle inclut le w_{max} calculé en c) il est possible que client ait utilisé en réalité une charge supérieure à w_{max} (une alarme). Si w_{max} est > la borne supérieure, alors il n'y a pas lieu de suspecter que w_{max} a été dépassé. Sur un grand nombre de tirs « n », on doit s'attendre à ce que $\alpha/2$ *n tirs génèrent des alarmes. Si le nombre observé est significativement supérieur à ce nombre, l'exécutant ne respecte probablement pas la norme fixée.

23- a) b_1 devrait être négatif, b_2 positif. b_1 est en $\%(\text{ohm-m})^{-1}$; b_2 est en %.

b) On observe une tendance linéaire dans chaque sous-groupe. Ceci indique que la régression ne présente probablement pas les mêmes coefficients pour les 2 types de roche. Un modèle permettant de tenir compte de cela est :

$$Cu = b_0 + b_1\rho_a + b_2 d + b_3 I_A + b_4\rho_a I_A + b_5 d I_A + e$$

où I_A est une indicatrice prenant la valeur 1 si la roche est de type a et 0 si elle est de type B.

c) $0.873^2=0.762$

d) $\dfrac{-0.285-(-0.873)*(-0.684)}{\left((1-0.873^2)*(1-0.684^2)\right)^{0.5}} = -0.877$, $R^2=0.762+(1-0.762)*(-0.877)^2=0.945$

---

24 a) test : F_{ajot}= ((10-8)/3) / (8/(10-6))=0.33 comparé à une $F_{table,\ 3,4,0.05}$=6.59. Clairement on ne peut rejeter H_0 donc il ne vaut pas la peine d'introduire la composante quadratique.

b) Comme il y a 10 données, il ne peut y avoir plus de 10 paramètres. On est donc limité à un polynôme d'ordre 3 qui compte exactement 10 termes. Il est dangereux d'utilser ce modèle car alors il ne reste plus aucun degré de liberté pour SCE. On ne peut donc pas juger si ce modèle est significatif. En fait on aura $R^2=1$ mais on aurait eu le même résultat avec tout ensemble de 10 variables, mêmes choisies aléatoirement.

c) $\sum_{i=1}^{9} u_i u_i^* = 100$ =SCR. Comme SCE vaut 10, SCT vaut 110.

---

25 a) oui. Avec 2 coefficients par exemple l'ellipse peut inclure la valeur (0,0) et celle-ci peut être hors de l'intervalle individuel sur chaque coefficient

b) oui. La valeur (0,0) peut être en dehors de l'ellipse de confiance et être à l'intérieur de chacun des intervalles individuels.

26- $Y=b_0+ b_1I_1 + b_2I_2+ b_3 X_1 +b_4I_1X_1 + b_5I_2X_1+e$

où I_1 est une indicatrice prenant la valeur 1 si l'observation provient du laboratoire 1, et 0 sinon.

I_2 est une indicatrice prenant la valeur 1 si l'observation provient du laboratoire 2, et 0 sinon.

l'ordonnée à l'origine pour labo 1 : b_1+b_1

l'ordonnée à l'origine pour labo 2 : b_0+b_2

l'ordonnée à l'origine pour labo 3 : b_0

La pente pour labo 1 est : b_3+b_4

la pente pour labo 2 est b_3+b_5

la pente pour labo 3 est : b_3

27- L'observation possède les valeurs ($\overline{y}, \overline{x}$), la moyenne des données sans la ième observation demeure donc ($\overline{y}, \overline{x}$), la droite de régression passera donc exactement à nouveau par ce point. La nouvelle droite de régression sera exactement la même que l'ancienne car l'ancienne minimise SCE et la contribution de l'observation « i » à l'ancien et au nouveau SCE est de 0. Si la nouvelle droite changeait ceci voudrait dire qu'on aurait pu en trouver une meilleure droite que celle trouvée avec l'observation « i » dans la régression. Comme la droite ne change pas, l'influence de l'observation est 0.

Autre justification : les deux droites passent par ($\overline{y}, \overline{x}$). La position de la droite au point x_i n'est donc pas modifiée. Par ailleurs, comme toutes les droites de régression passent par ($\overline{y}, \overline{x}$), ce point n'apporte aucune information sur la pente de la droite. Il n'y a donc aucune raison pour que le retrait de ce point influence la pente. La pente reste la même et la nouvelle droite passe par un même point ($\overline{y}, \overline{x}$) que l'ancienne, les deux droites sont donc confondues, et donc l'influence de l'observation « i » est 0.

28- On compare le F(ajout) successivement à des F(1,55), F(1,54), F(1,53), et F(1,52). Les valeurs sont approximativement 4.02. Conséquemment, on retient gamma et porosité seulement.

29- a)) b_1 : % $(\text{ohm-m})^{-1}$

c_1 : (ohm-m) $\%^{-1}$

b) Dans le cas de la régression à 1 variable X, on a $b_1=s_{xy}/s^2_x$ => $s_{xy}=b_1*s^2_x=0.5*7.6=3.8$

Pour l'autre régression on aura donc $c_1=3.8/10=0.38$

c) Modèle A car on veut prédire la concentration en sulfures à partir de la mesure de résistivité. Si l'on utilisait le modèle B, on aurait des prédictions beaucoup moins précises.
