Fiches Travaux Dirigés
Introduction à l'économétrie

CQLS
http://cqls.upmf-grenoble.fr

FICHE T.D. 1 Estimation des paramètres de régression et de bruit

Exercice 1 (Problème d'application sur un modèle de régression linéaire simple)
Cet exercice vise à conforter, via une étude économétrique, l'intuition suivante qui est que depuis 1961 les français n'ont cessé de substituer le cinéma par une consommation de films à domicile. Une idée consiste à mettre en relation la fréquentation des salles de cinéma avec le nombre de films TV. Pour répondre à cette problématique, on recueille deux séries `freq` et `films` observées annuellement de 1961 à 1998 (cf fin exercise). Ces séries représentent respectivement le nombre annuel d'entrées en millions et le nombre annuel de films diffuséés à la télévision.
On envisage un modèle linéaire expliquant la variable `log(freq)` en fonction de la seule variable `log(films)`. On tente une modélisation log-linéaire du type :

$$(log(freq))_i = \beta_0 + \beta_1 (log(films))_i + U_i, \qquad i = 1, \ldots, 38$$

Question 1 : Pourquoi les paramètres β_0 et β_1 ne sont pas calculables ?
Question 2 : On s'intéresse tout naturellement à l'estimation des paramètres β_0 et β_1. Déterminez les estimations obtenues par la méthode des moindres carrés.

On rappelle à titre indicatif que

- $var(\boldsymbol{x}) = \overline{x^2} - \overline{x}^2$
- $cov(\boldsymbol{x}, \boldsymbol{y}) = \overline{x \times y} - \overline{x} \times \overline{y}$.

```
> mean(log(films))
[1] 6.33244
> mean(log(films)^2)
[1] 40.67275
> mean(log(freq))
[1] 5.174373
> mean(log(freq)^2)
[1] 26.84506
> mean(log(films)*log(freq))
[1] 32.57717
```

Question 3 : Déterminez le coefficient de détermination linéaire (R^2) (puis le coefficient de corrélation linéaire (R)) entre `log(films)` et `log(freq)`, et donnez-en une interprétation.
Question 4 : Déterminez le coefficient de corrélation linéaire entre `log(films)` et `log(freq)` et donnez-en une interprétation.
Question 5 : Rappelez brièvement à quoi correspondent chacune des quatre colonnes de la matrice "Coefficients" de la sortie ci-dessous. Retrouvez les résultats des deux questions précédentes.

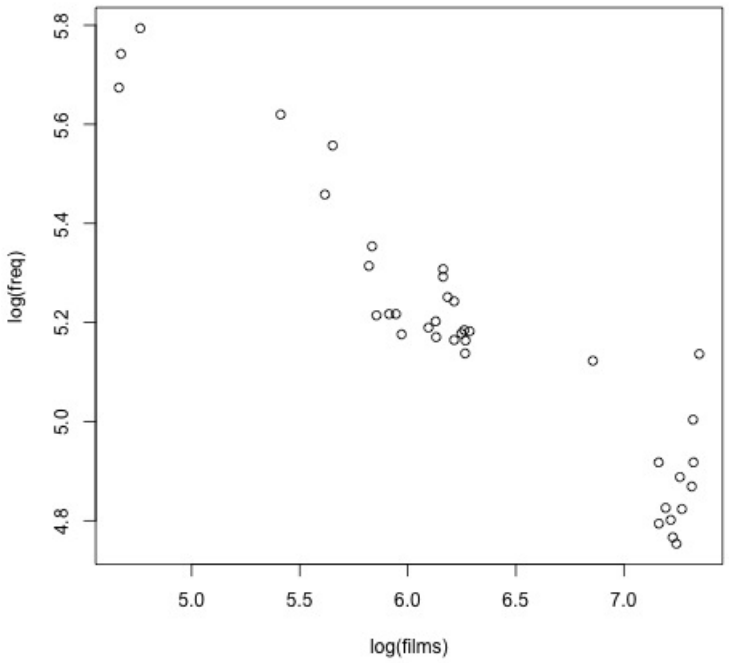
```
> summary(lm(log(freq)~log(films)))

Call:
lm(formula = log(freq) ~ log(films))

Residuals:
     Min       1Q   Median       3Q      Max
-0.12010 -0.06391 -0.01881  0.04304  0.29764

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.26592    0.12879   56.42   <2e-16 ***
log(films)  -0.33029    0.02019  -16.36   <2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09423 on 36 degrees of freedom
Multiple R-squared:  0.8814,        Adjusted R-squared:  0.8781
F-statistic: 267.5 on 1 and 36 DF,  p-value: < 2.2e-16

> sqrt(0.8813846)
[1] 0.9388209
```

Question 6 : Retrouvez les résultats des deux questions précédentes. Analysez brièvement les résultats obtenus.

```
> summary(lm(log(freq)~log(films)))

Call:
lm(formula = log(freq) ~ log(films))

Residuals:
     Min       1Q   Median       3Q      Max
-0.12010 -0.06391 -0.01881  0.04304  0.29764

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.26592    0.12879   56.42   <2e-16 ***
log(films)  -0.33029    0.02019  -16.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09423 on 36 degrees of freedom
Multiple R-squared:  0.8814,        Adjusted R-squared:  0.8781
F-statistic: 267.5 on 1 and 36 DF,  p-value: < 2.2e-16

> sqrt(0.8813846)
[1] 0.9388209
```

Question 7 : Sur le graphique ci-dessous reportez la droite ajustée (même approximativement) et illustrez la notion de valeur ajustée et de résidu.



Question 8 : En vous aidant de la sortie `R` précédente, quelle est l'estimation du niveau de bruit σ ?

Exercice 2 (consommation-revenu)

On considère le modèle économétrique très simple (et classique) liant la consommation et le revenu via une relation linéaire :

$$CONSO = \beta_0 + \beta_1 \times REVENU + U$$

Pour la suite, on supposera que le bruit U est centré et de variance σ^2. Les paramètres inconnus sont donc β_1, β_2 et σ^2. On souhaite ici obtenir les estimations de ces paramètres à partir du fichier `canada` constitué de $n = 35$ observations :

```
> canada
   REVENU  CONSO  T
1   30500  25747 61
2   33212  27576 62
3   35358  29417 63
4   38085  31421 64
...
32 619696 411068 92
33 631140 428143 93
34 644992 443982 94
35 666363 458083 95
```

Question 1 : (**Mode `R` calculette**)
Identifier les quantités calculées dans la sortie `R` suivante en utilisant les notations du cours et en rappelant les formules d'obtention.

```
> ## Indications R:
> ## %*% -> multiplication matricielle
> ## t() -> calcule la transposée
> ## solve() -> calcule l'inverse
> ## cbind() -> colle en colonnes des vecteurs ou matrices

> ## matrice de régression
```

```
> xx<-cbind(1,REVENU)
> ## estimations
> betaChapo<- solve(t(xx)%*%xx)%*%t(xx)%*%CONSO
> betaChapo
                [,1]
       3723.1088982
REVENU    0.6653596

> ## valeurs ajustées
> CONSOChapo<-xx%*%betaChapo

> ## résidus
> uChapo<-CONSO-CONSOChapo
> range(uChapo) ## range() -> min,max
[1] -14477.39  11107.30

> ## estimation du niveau (variance) de bruit
> sigma2Chapo<- sum(uChapo^2)/(length(uChapo)-2)
> sigma2Chapo
[1] 22777553
> sqrt(sigma2Chapo)
[1] 4772.584

> ## vecteurs des carrés des erreurs standard
> sigma2BetaChapo <- sigma2Chapo*diag(solve(t(xx)%*%xx))
> sigma2BetaChapo
                    REVENU
1.638408e+06 1.370262e-05
> sqrt(sigma2BetaChapo)
                    REVENU
1.280003e+03 3.701705e-03
```

Question 2 : (**Mode R utilisateur**) A partir de la sortie suivante, retrouver les estimations et leurs erreurs standard (estimations des écarts-type des estimateurs) des paramètres de régression ainsi que l'estimation du paramètre de bruit.

```
> summary(lm(CONSO~REVENU))

Call:
lm(formula = CONSO ~ REVENU)

Residuals:
     Min       1Q   Median       3Q      Max
-14477.4  -1322.4    713.9   2168.2  11107.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.723e+03  1.280e+03   2.909  0.00645 **
REVENU      6.654e-01  3.702e-03 179.744  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4773 on 33 degrees of freedom
Multiple R-squared:  0.999,        Adjusted R-squared:  0.9989
F-statistic: 3.231e+04 on 1 and 33 DF,  p-value: < 2.2e-16
```

FICHE T.D. 2 Tests d'hypothèses dans le cadre de la régression

Exercice 3 (suite exercice 2) On souhaite dans cet exercice vérifier les hypothèses de Keynes à partir d'un jeu de données :

1. la propension marginale à consommer est-elle inférieure à 1 ? On pourra répondre aux questions suivantes :
   - comment se décrit l'assertion d'intérêt en fonction du paramètre d'intérêt β_1 ? en fonction du paramètre d'écart $\delta_{\beta_1,1} := \frac{\beta_1 - 1}{\sigma_{\widehat{\beta_1}}}$?
   - quels sont les comportements aléatoires de $\widehat{\beta}_1(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ et $\widehat{\delta}_{\beta_1,1}(\boldsymbol{Y}|\underline{\boldsymbol{x}})$?
   - quelles sont les erreurs de décision que l'on peut commettre ? quelle est la plus grave de ces deux erreurs par rapport au rôle que l'on veut faire jouer aux données et pour quelle valeur du paramètre β_1 intervient-elle ? Idem pour le paramètre d'écart. (*Tentative d'indication* : Pour cela, nous pouvons imaginer le petit jeu d'argent suivant (qu'il faudrait en réalité envisager pour tous les tests que l'on met en place) : Avant que la réalité soit révélée, si le jeu de données semble plutôt valider l'assertion d'intérêt alors on mise une grosse somme d'argent sur l'assertion d'intérêt. Si en réalité, celle-ci se vérifie, on double sa mise et sinon on la perd.)
   - quelle est la pire des situations ? comment s'écrit-elle en fonction du paramètre d'intérêt ? en fonction du paramètre d'écart ?
   - peut-on construire la règle de décision à un seuil α en raisonnant sur le paramètre d'intérêt ? sur le paramètre d'écart ?
   - si vous avez répondu oui à l'une des deux questions précédentes (ce que nous espérons), calculez le quantile associé à la règle de décision pour un seuil de signification de 5% ?
   - nous sommes au jour J, quelles sont les estimations des paramètres et en particulier les valeurs de $\widehat{\beta}_1(\boldsymbol{y}|\underline{\boldsymbol{x}})$ et $\widehat{\delta}_{\beta_1,1}(\boldsymbol{y}|\underline{\boldsymbol{x}})$? Appliquez alors la règle de décision (voir indication ci-dessous).
   - quelle est la règle de décision alternative basée sur l'indicateur du niveau de fiabilité de la règle de décision ?
   - rédigez le test sous une forme standard.
2. La propension marginale à consommer est-elle strictement positive au seuil de 5% ?
3. Sans calcul suppléméntaire montrez que la propension marginale est différente de zéro au seuil de 5%. Ce test porte le nom plus connu de test de significativité locale car il permet de montrer que le régresseur `REVENU` apporte une information significative dans l'explication de la `CONSO`.
4. La consommation incompressible est-elle strictement positive (toujours au seuil de 5%) ?
5. Un néophyte a souhaité effectuer un test au seuil de 5% pour essayer de savoir si $\beta_1 > 0.7$ et a obtenu avec ses données et son logiciel préféré une p-valeur de l'ordre de 99.2%.
   - Proposez l'instruction R qui permet d'obtenir la valeur de cette p−valeur.
   - Comment doit-il conclure au test qu'il a mis en place ?
   - Peut-il conclure autre chose ?
   - Quel est le risque mimimum qu'il doit prendre s'il souhaite montrer que $\beta_1 \neq 0.7$?

Indication :

```
> summary(lm(CONSO~REVENU))

Call:
lm(formula = CONSO ~ REVENU)

Residuals:
     Min       1Q   Median       3Q      Max
-14477.4  -1322.4    713.9   2168.2  11107.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.723e+03  1.280e+03   2.909  0.00645 **
REVENU      6.654e-01  3.702e-03 179.744  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4773 on 33 degrees of freedom
Multiple R-squared:  0.999,        Adjusted R-squared:  0.9989
F-statistic: 3.231e+04 on 1 and 33 DF,  p-value: < 2.2e-16

> (0.6654-1)/0.003702
[1] -90.38358
> pt(-90.38358,35-2)
[1] 2.06455e-41
> pnorm(-90.38358)
[1] 0
```

Exercice 4 (consommation de champage) On souhaite étudier l'influence du revenu personnel (variable `R`), du prix du champagne (variable `P`)et du prix des liqueurs (variable `PL`) sur la consommation de champagne (variable `C`). Un spécialiste pense qu'un modèle adapté est le modèle log-linéaire suivant :

$$\log(C) = \beta_0 + \beta_1 \log(R) + \beta_2 \log(P) + \beta_3 \log(PL) + U$$

On supposera que le bruit U est centré et de variance σ^2.

1. Ce modèle est-il très loin du modèle linéaire (traité depuis le début de l'année) ?
2. Quelles sont les estimations des paramètres basés sur le jeu de données `champ` ?
3. Complétez : lorsque le revenu _ _ _ _ _ _ _ de 10%, la consommation de champagne _ _ _ _ _ _ _ approximativement de _ _ _ _ _ _ _ .
4. Peut-on penser que le régresseur prix du champagne (`P`) apporte de l'information dans l'explication de la consommation de champagne, i.e. est-il significatif ? (*Indication* : conditions mathématiques d'utilisation à préciser si nécessaire)
5. Même question pour le régresseur revenu (`R`).
6. Même question pour le régresseur prix des liqueurs (`PL`).
7. Peut-on penser au seuil de 5% que le champagne est un produit de luxe ?

```
> champ
      C     R      P    PL
1  42.5  57.2  76.60  73.6
2  38.7  59.1  80.70  72.9
3  40.0  61.5  86.80  67.0
4  45.4  64.0  85.40  63.2
5  51.7  67.6  84.10  55.1
6  65.4  71.7  81.70  59.2
7  72.4  75.5  80.90  65.6
```

```
8   59.0  76.2  91.70  59.5
9   61.3  78.0  96.50  56.8
10  75.6  81.9  95.40  61.3
11  82.5  86.6  96.20  73.0
12  90.5  93.0  98.41  88.9
13 100.0 100.0 100.00 100.0
14 110.5 105.4  99.30 111.4
15 127.9 109.8  97.80 123.3
16 139.0 115.0  96.00 136.9
17 143.8 120.5 104.30 150.6
> summary(lm(log(C)~log(R)+log(P)+log(PL),data=champ))

Call:
lm(formula = log(C) ~ log(R) + log(P) + log(PL), data = champ)

Residuals:
      Min        1Q    Median        3Q       Max
-0.066608 -0.031357 -0.006315  0.022400  0.067713

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.48769    0.74597   0.654   0.5247
log(R)       2.36686    0.13037  18.155 1.28e-10 ***
log(P)      -1.36409    0.24213  -5.634 8.15e-05 ***
log(PL)     -0.10679    0.06013  -1.776   0.0991 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04215 on 13 degrees of freedom
Multiple R-squared:  0.9923,        Adjusted R-squared:  0.9906
F-statistic: 560.5 on 3 and 13 DF,  p-value: 5.408e-14

> qt(.975,17-3-1)
[1] 2.160369
> qnorm(.975)
[1] 1.959964
> ####################
> 1-pt( (2.36686-1)/0.13037,13 )
[1] 5.181925e-08
```

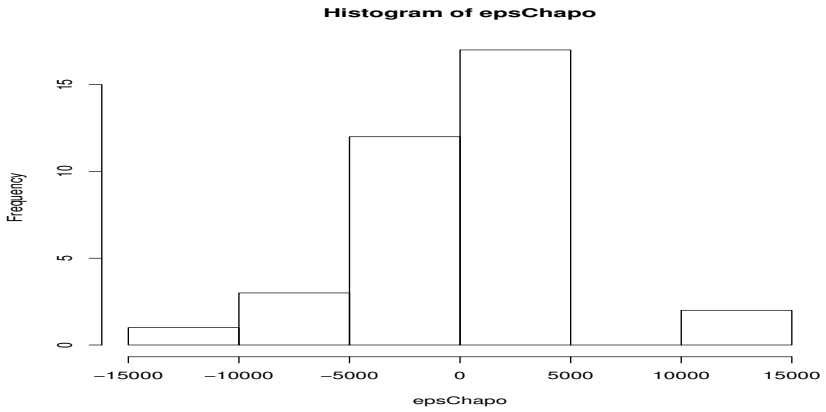
Exercice 5 (suite des exercices 2 et 3)

1. On se place dans le cadre asymptotique (n grand). Peut-on penser au vu des données que le paramètre de nuisance $\sigma > 3900$?

2. Que se serait-il passé si on avait fait l'hypothèse que le bruit est gaussien, que devient le test précédent ?

```
> sigma2Chapo
[1] 22777553
> sqrt(sigma2Chapo)
[1] 4772.584
> ####################
> (sigma2Chapo-3900^ 2)/sqrt(var(uChapo^ 2)/35)
         [,1]
[1,] 1.017061
> 1-pnorm(1.017061)
[1] 0.1545622
> ####################
> qchisq(0.95,35-2)
```

```
[1] 47.39988
> (35-2)*sigma2Chapo/3900^ 2
[1] 49.41875
> 1-pchisq(49.41875,33)
[1] 0.03304546
> ##################  Annexe
> shapiro.test(uChapo)

        Shapiro-Wilk normality test

data:  uChapo
W = 0.9123, p-value = 0.008634
```



Exercice 6 (suite exercice 1)

7. Quelles sont les hypothèses de test qui permettrait de montrer que le régresseur `log(films)` apporte de l'information dans l'explication de la fréquentation des salles de cinéma ?

8. En analysant la précédente sortie `R`, pourriez-vous justifier (au vu des données) le fait que l'action de la variable `log(films)` dans l'explication de la variable `log(freq)` est très significative.

9. (difficile) Donnez un ordre de grandeur de la p−valeur du test

$$H_0 \ : \ \beta_1 = 0 \quad \text{contre} \quad H_1 \ : \ \beta_1 < 0.$$

Au seuil de 1%, peut-on affirmer qu'une certaine augmentation en pourcentage du nombre de films diffusés à la TV entraînerait une **diminution** d'un certain pourcentage de la fréquentation des salles de cinéma ? (Indication : représentez la p−valeur (très approximativement puisqu'elle est très faible) du test $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ sur le graphique ci-dessous représentant une $\mathcal{S}t(38 - 1 - 1)$)

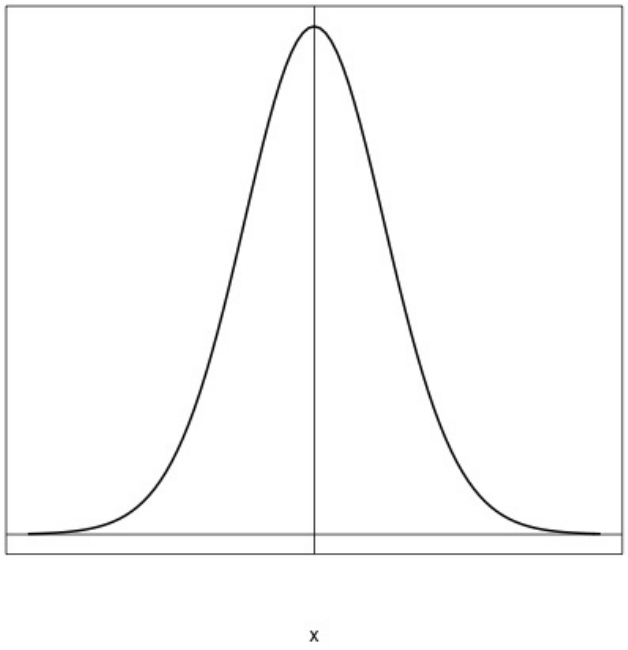


Figure 2.1: Densité d'une loi de $\mathcal{S}t(36)$

FICHE T.D. 3 Phénomène de colinéarité

Exercice 7 (Colinéarité "détectable par matrice de corrélation")

Considérons le jeu de données pédagogique `colinEx`. Ce jeu de données de taille 100 décrit quatre variables totalement fictives : une variable à expliquer `y` et trois régresseurs quantitatifs `x1`, `x2` et `x3`.

Question 1 : A partir des traitements préliminaires ci-dessous quels commentaires êtes-vous amenés à faire ?

```
> summary(lm(y~x1+x2+x3))

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89993 -0.70514  0.06399  0.76046  2.10918

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7173     0.2640   2.717  0.00782 **
x1            7.7565     9.3562   0.829  0.40915
x2            3.4461     0.3808   9.051 1.63e-14 ***
x3           -1.5123     9.3557  -0.162  0.87193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.027 on 96 degrees of freedom
Multiple R-squared:  0.8149,        Adjusted R-squared:  0.8091
F-statistic: 140.8 on 3 and 96 DF,  p-value: < 2.2e-16
```

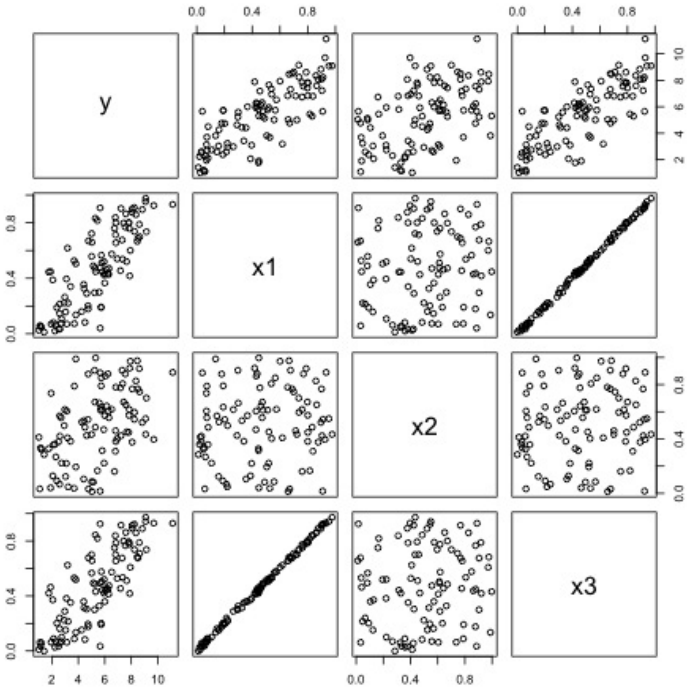
Question 2 : Un praticien non expérimenté au vu des résultats précédents poursuit son analyse de la manière suivante. Qu'en pensez-vous ?

```
> summary(lm(y~x2))

Call:
lm(formula = y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4023 -1.5064  0.0714  1.5932  4.7542

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3093     0.4389   7.540 2.39e-11 ***
x2            4.1315     0.7646   5.403 4.59e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.073 on 98 degrees of freedom
Multiple R-squared:  0.2295,        Adjusted R-squared:  0.2217
F-statistic:  29.2 on 1 and 98 DF,  p-value: 4.594e-07
```

Analysez en particulier la chute du coeffcient de détermination.

Question 3 : Particulièrement surpris par ces résultats, il demande conseil à un de ses collègues qui lui conseille de calculer la matrice de corrélation et d'afficher tous les nuages de points croisant les deux variables.

```
> cor(colinEx)
           y        x1        x2        x3
y  1.0000000 0.8104127 0.4790971 0.8094144
x1 0.8104127 1.0000000 0.1032505 0.9992734
x2 0.4790971 0.1032505 1.0000000 0.1028280
x3 0.8094144 0.9992734 0.1028280 1.0000000
```



Au vu de ces résultats, il s'empresse alors de lancer les deux régressions simples suivantes apparemment rejetées par sa première analyse. Quelle conclusion peut-on en tirer ?

```
> summary(lm(y~x1))

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4622 -0.8642  0.0737  0.9347  3.0870

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3006     0.2652   8.673 9.09e-14 ***
x1            6.5803     0.4805  13.694  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.383 on 98 degrees of freedom
Multiple R-squared:  0.6568,        Adjusted R-squared:  0.6533
F-statistic: 187.5 on 1 and 98 DF,  p-value: < 2.2e-16

> summary(lm(y~x3,data=colinEx))
```

```
22
23 Call:
24 lm(formula = y ~ x3, data = colinEx)
25
26 Residuals:
27     Min      1Q  Median      3Q     Max
28 -3.4386 -0.8432  0.1037  0.9589  3.1305
29
30 Coefficients:
31             Estimate Std. Error t value Pr(>|t|)
32 (Intercept)   2.2960     0.2664   8.619 1.19e-13 ***
33 x3            6.5721     0.4817  13.645  < 2e-16 ***
34 ---
35 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
36
37 Residual standard error: 1.387 on 98 degrees of freedom
38 Multiple R-squared:  0.6552,        Adjusted R-squared:  0.6516
39 F-statistic: 186.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

Question 4 : Son collègue lui rappelle alors une règle d'or : il est dangereux dans une sélection de modèle pas à pas de retirer (ou ajouter) plus d'un régresseur. Il applique alors cette règle en repartant de sa première analyse et en ne retirant que le regresseur étant le moins significatif.

```
1  > summary(lm(y~x1+x2))
2
3  Call:
4  lm(formula = y ~ x1 + x2)
5
6  Residuals:
7      Min      1Q  Median      3Q     Max
8  -2.8599 -0.7236  0.0752  0.7499  2.1256
9
10 Coefficients:
11             Estimate Std. Error t value Pr(>|t|)
12 (Intercept)   0.7146     0.2622   2.726  0.00762 **
13 x1            6.2452     0.3567  17.509  < 2e-16 ***
14 x2            3.4467     0.3788   9.098 1.19e-14 ***
15 ---
16 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
17
18 Residual standard error: 1.021 on 97 degrees of freedom
19 Multiple R-squared:  0.8148,        Adjusted R-squared:  0.811
20 F-statistic: 213.4 on 2 and 97 DF,  p-value: < 2.2e-16
```

Question 5 : Il demande alors à son collègue s'il connaît la raison de tels résultats. Ce dernier lui indique alors qu'une colinéarité "forte" des régresseurs peut engendrer une très forte variabilité des estimateurs et ainsi une difficulté à rejeter l'hypothèse de significativité de certains paramètres de régression (même ceux "très" corrélés à la variable à expliquer). A partir de la matrice de corrélation, détecter une forte corrélation entre certains régresseurs et à partir des calculs matriciels ci-dessous (où $A\% * \%B$, $solve(A)$ et $t(A)$ calcule respectivement le produit matriciel entre A et B, l'inverse et la transposée de la matrice A) mettre en évidence le changement des variabilité des estimateurs.

```
1  > x<-cbind(1,x1,x2,x3)
2  > solve(t(x)%*%x)
3                          x1          x2           x3
4      0.06615065   0.09716959 -0.06335320  -0.14793913
5  x1  0.09716959  83.06832374 -0.04433339 -83.00315927
6  x2 -0.06335320  -0.04433339  0.13757078   0.03098143
7  x3 -0.14793913 -83.00315927  0.03098143  83.05998690
```

```
> solve(t(x[,-4])%*%x[,-4])
                          x1          x2
    0.06588716 -0.05066832 -0.06329802
x1 -0.05066832  0.12195321 -0.01337315
x2 -0.06329802 -0.01337315  0.13755923
```

Exercice 8 (Mise en évidence de phénomène de colinéarité)

L'idée est de considérer un modèle théorique dont les régresseurs possèdent un "certain" niveau de colinéarité entre eux, ceci afin d'appréhender les conséquences que cela peut engendrer sur les comportements des estimateurs des paramètres du modèle de régression.

On se propose d'étudier des modèles de régression linéaire à trois régresseurs satisfaisant les hypothèses classiques. Étant donné un vecteur $\boldsymbol{\beta}$, et une taille d'échantillon n, on définit trois vecteurs indépendants, réalisations d'une loi uniforme sur $[0,1]$, notés $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$. Le modèle considéré s'écrit alors :

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3' + \boldsymbol{U},$$

où l'on définit $\boldsymbol{x}_3'$ comme une "quasi"-combinaison linéaire de $\boldsymbol{x}_1, \boldsymbol{x}_2$ et $\boldsymbol{x}_3$:

$$\boldsymbol{x}_3' = \alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2 + (1 - \alpha_1 - \alpha_2)\boldsymbol{x}_3 + \boldsymbol{U}'.$$

Le vecteur $\boldsymbol{U}'$ constitue une petite perturbation de la combinaison linéaire "pure"qui permet que la matrice $(\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3')$ soit inversible!

Le vecteur $\boldsymbol{\beta}$, la taille d'échantillon n ainsi que l'écart-type σ relatif bruit $\boldsymbol{U}$ sont fixés comme ci-dessous :

$$\boldsymbol{\beta} = (1,2,3,4)^T, n = 200, \sigma = .2$$

On donne quatre vecteurs $\boldsymbol{\alpha}^A, \boldsymbol{\alpha}^B, \boldsymbol{\alpha}^C, \boldsymbol{\alpha}^D$ définissant 4 modèles (A, B, C et D) distincts de colinéarité.

$$\boldsymbol{\alpha}^A = (0.05, 0.95, 0), \quad \boldsymbol{\alpha}^B = (0,0,1), \quad \boldsymbol{\alpha}^C = (0.5, 0.5, 0) \quad \text{et} \quad \boldsymbol{\alpha}^D = (0.95, 0.05, 0).$$

Question 1 : Pour chacun des quatres modèles, exprimez le niveau de colinéarité entre les régresseurs : absence de colinéarité, forte colinéarité,...

Question 2 : On observe quatre matrices de corrélation M_1, M_2, M_3 et M_4 correspondant chacune à un modèle (A,B,C,D), quelles associations peut-on espérer?

```
> M1
                x1         x2    xPrime3
x1      1.00000000 0.06253862 0.96014670
x2      0.06253862 1.00000000 0.08805359
xPrime3 0.96014670 0.08805359 1.00000000
> M2
                x1         x2   xPrime3
x1      1.00000000 0.06253862 0.7060393
x2      0.06253862 1.00000000 0.6505921
xPrime3 0.70603934 0.65059208 1.0000000
> M3
                 x1          x2     xPrime3
x1       1.00000000  0.06253862 -0.10538403
x2       0.06253862  1.00000000 -0.03339283
xPrime3 -0.10538403 -0.03339283  1.00000000
> M4
                x1         x2   xPrime3
x1      1.00000000 0.06253862 0.1256600
x2      0.06253862 1.00000000 0.9539574
xPrime3 0.12566000 0.95395738 1.0000000
```

Question 3 : Pensez-vous pouvoir à la seule vue des matrices de corrélation pouvoir détecter tous les types de colinéarité ?

Question 4 : Rappelez d'où vient la définition du VIF (Variance Inflation Factor). En analysant le calcul des vif de chaque modèle dans l'analyse, et sachant que le VIF_1 (resp. VIF_2, VIF_3 et VIF_4) est calculé avec les régresseurs intervenant dans M_1 (resp. M_2, M_3 et M_4), retrouvez les associations faites à la question précédente.

```
> VIF1
       x1        x2   xPrime3
 1.058945 11.582197 11.735832
> VIF2
      x1       x2  xPrime3
4.417289 3.822700 7.657322
> VIF3
       x1        x2   xPrime3
12.989023  1.015794 13.045355
> VIF4
      x1       x2  xPrime3
1.014705 1.005321 1.010833
```

Question 5 : L'analyse du vif permet-elle de mieux appréhender la colinéarité entre régresseurs ?

Exercice 9

La problématique est de savoir quels sont les facteurs influençant l'utilisation mondiale d'internet (via le nombre d'internautes). L'analyse sera basée sur un **jeu de données décrit à la fin de l'exercise**. Dans un premier temps, on va regarder l'influence de la taille de la population et du nombre d'ordinateurs individuels pour 37 pays. On tente alors une modélisation de type log-linéaire sous la forme

$$(\log(int))_i = \beta_0 + \beta_1(\log(ord))_i + \beta_2(\log(pop))_i + U_i, \quad i = 1, \ldots, 37.$$

Rappelons qu'envisager un tel modèle revient à supposer que l'élasticité de *ord* sur *int* et de *pop* sur *int* sont **constantes**. Par la suite, on supposera (sans y prêter une quelconque attention) que le bruit $\boldsymbol{U}$ satisfait les hypothèses classiques des modèles linéaires (présentées dans le polycopié de cours).

Question 1 : Analysez les sorties ci-dessous

```
> summary(lm(log(int)~log(ord)+log(pop))->reg)

Call:
lm(formula = log(int) ~ log(ord) + log(pop))

Residuals:
     Min       1Q   Median       3Q      Max
-0.96836 -0.38059 -0.01449  0.21075  2.02775

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.31267    0.79065   1.660 0.106063
log(ord)     0.66784    0.06644  10.052 1.02e-11 ***
log(pop)     0.30136    0.06877   4.382 0.000107 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.628 on 34 degrees of freedom
Multiple R-squared:  0.7913,        Adjusted R-squared:  0.779
F-statistic: 64.45 on 2 and 34 DF,  p-value: 2.708e-12

> vif(reg)
log(ord) log(pop)
1.007644 1.007644
> 1-1/vif(reg)
   log(ord)    log(pop)
0.007586332 0.007586332
```

Question 2 : On envisage alors en plus d'indicateurs descriptifs d'intégrer un indicateur économique à savoir le `pib` de chaque pays. Au vu de la sortie ci-dessous écrire l'équation du modèle et interprétez les résultats. Que confirme la figure Fig. 3.1 ?

```
> summary(lm(log(int)~log(ord)+log(pop)+log(pib))->reg2)

Call:
lm(formula = log(int) ~ log(ord) + log(pop) + log(pib))

Residuals:
     Min       1Q   Median       3Q      Max
-0.93074 -0.26083 -0.09216  0.25388  1.05011

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.82328    0.67975   4.153 0.000217 ***
log(ord)     0.20389    0.10662   1.912 0.064538 .
log(pop)    -0.03787    0.08647  -0.438 0.664242
log(pib)     0.85683    0.17286   4.957 2.09e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4826 on 33 degrees of freedom
Multiple R-squared:  0.8804,        Adjusted R-squared:  0.8695
F-statistic: 80.94 on 3 and 33 DF,  p-value: 2.691e-15

> vif(reg2)
log(ord) log(pop) log(pib)
4.393758 2.697517 6.492686
> 1-1/vif(reg2)
 log(ord)  log(pop)  log(pib)
0.7724044 0.6292887 0.8459805
```

Question 3 : Quelle règle de conduite a été adoptée pour obtenir la sortie ci-dessous ? Interprétez.

```
> summary(lm(log(int)~log(ord)+log(pib))->reg3)

Call:
lm(formula = log(int) ~ log(ord) + log(pib))

Residuals:
    Min      1Q  Median      3Q     Max
-0.9242 -0.2619 -0.1007  0.2438  1.0186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.58132    0.39136   6.596 1.47e-07 ***
log(ord)     0.23515    0.07827   3.005  0.00497 **
log(pib)     0.79690    0.10438   7.634 7.15e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4768 on 34 degrees of freedom
Multiple R-squared:  0.8797,        Adjusted R-squared:  0.8726
F-statistic: 124.3 on 2 and 34 DF,  p-value: 2.327e-16

> vif(reg3)
log(ord) log(pib)
2.425311 2.425311
> 1-1/vif(reg3)
 log(ord)  log(pib)
0.5876818 0.5876818
```

Question 4 : a) Comment retrouver la p−valeur du test de significativité locale du régresseur `log(pib)` (indication : on pourra, pour ceux qui le souhaitent, répondre en fournissant l'instruction `R` permettant de la calculer).

b) A l'aide d'une calculatrice (ou par simple calcul mental), peut-on penser au vu des données que $\beta_2 < 1$ au seuil de 5% (remarquez que n est suffisamment grand pour ...) ? Faites un dessin (à main levée) représentant la règle de décision et la p−valeur de ce test.

c) (**question pour les experts**) Sans AUCUN CALCUL, proposez une bonne approximation de la p−valeur du test $H_1 : \beta_2 < 1.6$ (indication : $2 \times 0.7969 \simeq 1.6$).

```
> internet
            pays        int         pop       ord  pib
1    etats-unis 123326.000  278058.881 49896.200 7746
2         japon  63955.200  126771.662  7485.780 4202
3     allemagne  28876.900   83029.536  3914.060 2100
...
33        chine  28697.200 1280775.530   140.599  996
34    thailande   1567.700   61797.751    66.000  157
35     colombie    634.055   40349.388    52.160   85
36         inde   4748.760 1029991.150    45.420  360
37  philippines    410.127   82841.518    29.460   83
```

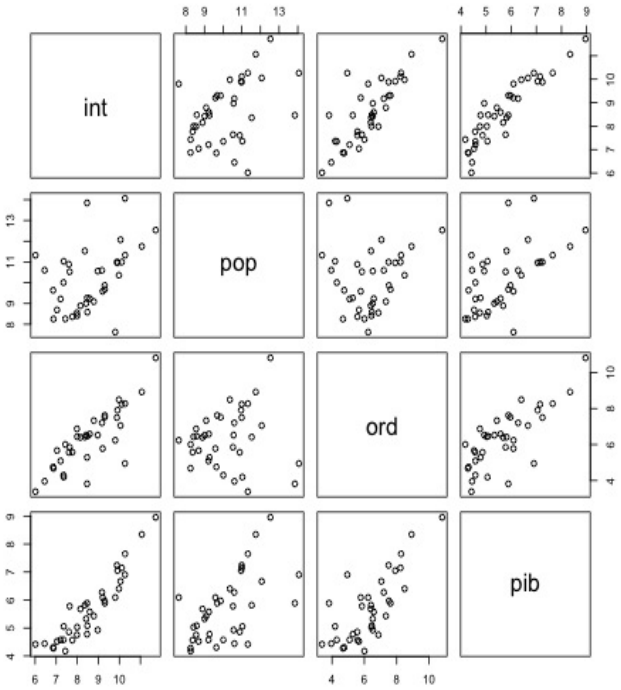


Figure 3.1: Plots deux à deux des variables du jeu de données internet transformées logarithmiquement.

FICHE T.D. 4 Régression avec Covariables qualitatives

Exercice 10

Trois praticiens sont désireux d'étudier le salaire (noté Sal) d'un individu en fonction d'un indice sur son niveau d'expérience professionnelle (noté Exp) (compris entre 0 et 1 et créé par de brillants spécialistes). Cette étude se limite à des individus dans un certain secteur d'activité. Une variable $IndExp$ a été introduite pour fabriquer des groupes de niveau d'expérience à partir de la variable Exp :

$$IndExp = \begin{cases} Bas & \text{si } 0 \leq Exp < \frac{1}{3} \\ Moyen & \text{si } \frac{1}{3} \leq Exp < \frac{2}{3} \\ Haut & \text{si } \frac{2}{3} \leq Exp \leq 1 \end{cases}$$

Enfin, ils disposent aussi de la variable Sex (1=Homme et 0=Femme). Pour se divertir, ils décident que chacun d'entre eux propose leur propre traitement à partir d'un même jeu de données :

- la taille d'échantillon est de n=300 individus
- il y a autant (i.e. 50) d'individus de l'échantillon dans chacune des six catégories possibles en croisant les deux variables qualitatives Sex et $IndExp$.

Question 1 : *Traitement du premier praticien* : Appréciant les traitements simples, il tente une régression simple de Sal en fonction de Exp dont le résumé est fourni par la commande `R` suivante :

```
> summary(lm(Sal~Exp))

Call:
lm(formula = Sal ~ Exp)

Residuals:
    Min      1Q  Median      3Q     Max
-859.44 -282.47    9.84  297.94  821.64

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   866.44      42.05   20.61   <2e-16 ***
Exp          1815.91      71.99   25.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 366.3 on 298 degrees of freedom
Multiple R-squared:  0.681,        Adjusted R-squared:   0.68
F-statistic: 636.3 on 1 and 298 DF,  p-value: < 2.2e-16
```

Traitement du deuxième praticien : Ce praticien plus expérimenté connaît mieux les modèles de régression. Il décide de traiter le modèle ci-dessous faisant intervenir à la fois la variable a priori qualitative Sex (pouvant être considéré comme une variable quantitative) et l'indice Exp d'expérience professionnelle :

$$Sal_i = \beta_0 + \beta_1 Exp_i + \beta_2 Sex_i + \beta_3 (Exp_i \times Sex_i) + U_i$$

```
>  summary(lm(Sal~Exp*Sex))

Call:
lm(formula = Sal ~ Exp * Sex)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-478.06 -142.38   -9.41  143.13  595.07

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   756.96      32.35  23.400  < 2e-16 ***
Exp          1448.66      57.48  25.201  < 2e-16 ***
Sex           294.45      46.98   6.267 1.29e-09 ***
Exp:Sex       575.75      80.46   7.156 6.57e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 204.1 on 296 degrees of freedom
Multiple R-squared:  0.9016,        Adjusted R-squared:  0.9007
F-statistic: 904.6 on 3 and 296 DF,  p-value: < 2.2e-16
```

Question : *Comparer le coefficient de détermination R^2. Pour les deux modèles, quelle(s) valeur(s) prédiriez-vous pour les salaires d'une femme et d'un homme ayant un indice d'expérience professionnelle égal à* 0.5 *?*

Question 2 : *Traitement du troisième praticien* : Spécialisé dans les modèles ANOVA (moins dans les modèles de régression), il se propose d'expliquer le Salaire (*Sal*) en fonction des facteurs (variables qualitatives) *Sex* et *IndExp*. L'instruction `R` (avec sa sortie standard) conduisant à une telle analyse est donnée ci-dessous :

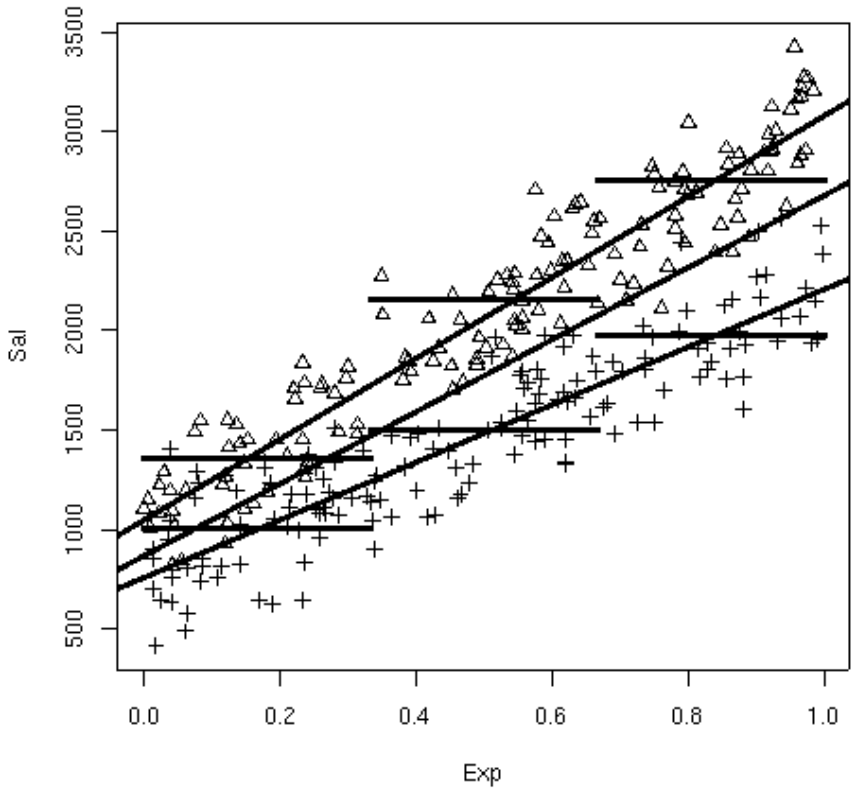
```
> summary(aov(Sal~IndExp*Sex))
             Df   Sum Sq  Mean Sq F value   Pr(>F)
IndExp        2 74368397 37184199  493.40  < 2e-16 ***
Sex           1 26370087 26370087  349.91  < 2e-16 ***
IndExp:Sex    2  2463488  1231744   16.34 1.86e-07 ***
Residuals   294 22156811    75363
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Afin de comparer ses résultats avec ceux de ses collègues, ce praticien (malgré sa moins bonne connaissance des modèles de régression) sait toutefois qu'un modèle ANOVA peut s'écrire comme un modèle de régression linéaire et propose l'analyse suivante :

```
> summary(lm(Sal~IndExp*Sex))

Call:
lm(formula = Sal ~ IndExp * Sex)

Residuals:
    Min      1Q  Median      3Q     Max
-643.05 -193.87   -3.37  177.69  667.47

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1006.97      37.36  26.955  < 2e-16 ***
IndExpMoyen       490.07      52.59   9.318  < 2e-16 ***
IndExpHaut        965.24      56.11  17.203  < 2e-16 ***
Sex               348.07      55.08   6.319 9.72e-10 ***
IndExpMoyen:Sex   309.93      76.89   4.031 7.08e-05 ***
IndExpHaut:Sex    436.40      78.97   5.526 7.21e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 274.5 on 294 degrees of freedom
Multiple R-squared:  0.8233,        Adjusted R-squared:  0.8202
F-statistic: 273.9 on 5 and 294 DF,  p-value: < 2.2e-16
```

Question 3 : A partir de la dernière sortie, écrire l'équation du modèle (*Indication :* en `R`, `IndExpMoyen` correspond à l'indicatrice associée à l'événement $IndExp = Moyen$).
Question 4 : Quelle(s) valeur(s) prédiriez-vous pour les salaires d'une femme et d'un homme ayant un indice d'expérience professionnelle égal à 0.5 ?
Question 5 : *Conclusions* : Les trois praticiens confrontent enfin leurs résultats.
Question 6 : Quel(s) modèle(s) vous semble(nt) être le(s) plus intéressant(s) pour le but que se sont fixés les trois praticiens ? (Justifier votre réponse)

Pour comparer les résultats, ils proposent de représenter sur un même graphique les différents modèles ajustés ainsi que le nuage des individus (Triangle=Homme et Croix=Femme) :



Question 7 : Identifier sur le graphique les droites et segments de droites représentant les trois modèles proposés par ces praticiens.

Exercice 11

Un praticien se propose d'expliquer le salaire d'une certaine catégorie de la population active en fonction du niveau d'étude et de l'expérience professionnelle. Il s'appuie pour cette analyse sur un jeu de données recueilli auprès de $n = 200$ individus constitué du salaire mensuel (variable `Sal`), d'un indicateur du niveau d'études (variable `IndEtu`) et d'un indicateur de l'expérience professionnelle (variable `IndExp`). Ces deux indicateurs ont été calculés par un expert de sorte qu'ils varient entre 0 et 1.

Question 1 : On envisage un modèle linéaire standard

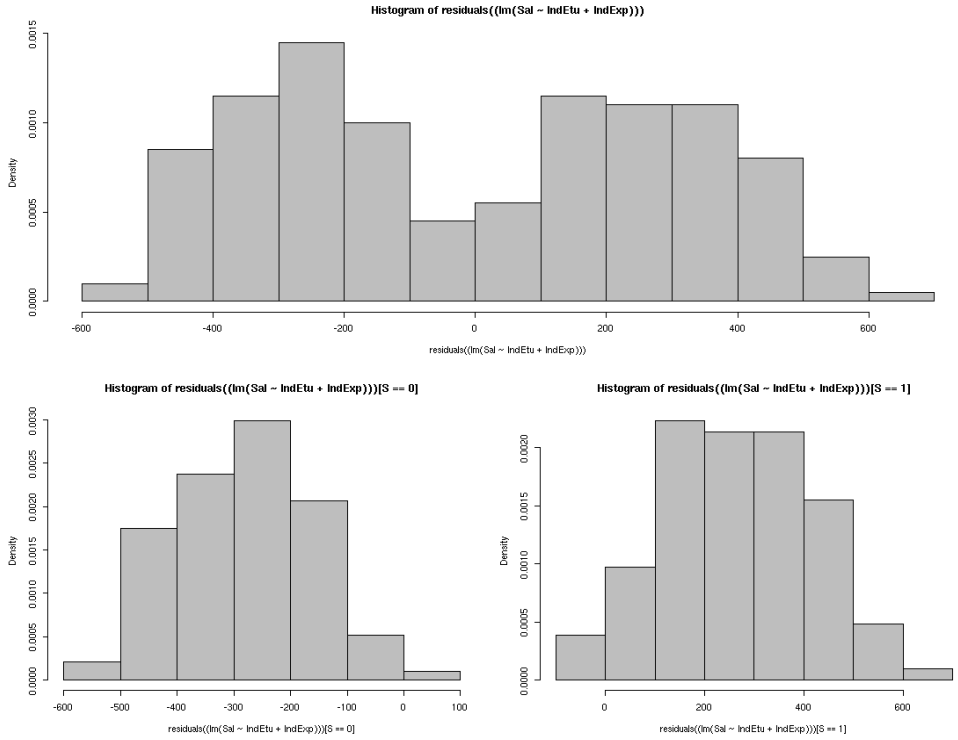
$$\mathbf{Sal} = \beta_0 + \beta_1\, \mathbf{IndEtu} + \beta_2\, \mathbf{IndExp} + \boldsymbol{U}.$$

Analysez les résultats de la régression présentés dans le tableau ci-après.

```
> summary(lm(Sal~IndEtu+IndExp))

Call:
lm(formula = Sal ~ IndEtu + IndExp)

```

```
Residuals:
    Min      1Q  Median      3Q     Max
-539.07 -265.83    1.85  273.36  644.93

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   928.93      55.56  16.720  < 2e-16 ***
IndEtu        238.91      71.25   3.353 0.000959 ***
IndExp       1489.23      73.94  20.140  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 309.8 on 197 degrees of freedom
Multiple R-squared:  0.6787,        Adjusted R-squared:  0.6755
F-statistic: 208.1 on 2 and 197 DF,  p-value: < 2.2e-16
```

Question 2 : N'ayant pas les moyens d'observer le nuage de points, le praticien sait qu'un bon réflexe est d'observer la répartition des résidus (premier histogramme en haut du graphique ci-après). Que remarquez-vous?



Question 3 : Le praticien décide alors d'introduire une variable binaire S (égale à 1 ou 0) permettant ainsi de classer les individus en deux catégories. Après avoir récolté les informations supplémentaires relatives à S, il "plot" les deux autres histogrammes ci-dessus représentant les répartitions des résidus par catégorie. Pourriez-vous expliquer l'analyse du praticien? Et compte tenu de la problématique, quel vous semble être la nature de cette variable discriminante S?

Question 4 : Fort de cette interprétation graphique, le praticien décide d'intégrer S dans le modèle. Cependant, il manque d'expérience dans le traitement de ce type de problème. Lorsqu'il y a deux régresseurs quantitatifs, il sait que la méthode MCO consiste à déterminer dans l'espace de représentation porté par les trois variables le plan le plus proche ("verticalement") du nuage de points. En revanche, il ne visualise pas très bien ce que fera la méthode après introduction de la variable S. Dans cet espace de représentation (avec les trois mêmes variables), pourriez-vous lui expliquer comment s'interprète la méthode MCO?

Question 5 : Après s'être informé, le praticien envisage alors deux nouveaux modèles :

→ modèle A : on ajoute $\boldsymbol{S}$ au modèle initial.

→ modèle B : on ajoute $\boldsymbol{S} \times \boldsymbol{IndEtu}$ (notée `S:IndEtu` en `R`) et $\boldsymbol{S} \times \boldsymbol{IndExp}$ (notée `S:IndExp` en `R`) au modèle A.

Exprimez géométriquement (via les caractéristiques des plans associés aux deux catégories) la différence entre les deux nouveaux modèles?

Question 6 : Déterminez l'équation de chaque modèle, analysez les résultats de chaque régression et comparez-les avec ceux du modèle initial.

Modèle A :

```
> summary(lm(Sal~IndEtu+IndExp+S))

Call:
lm(formula = Sal ~ IndEtu + IndExp + S)

Residuals:
    Min      1Q  Median      3Q     Max
-309.08 -104.64    6.17  101.56  364.73

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   621.74      27.22  22.845  < 2e-16 ***
IndEtu        250.43      31.95   7.839 2.83e-13 ***
IndExp       1525.13      33.17  45.976  < 2e-16 ***
S             550.78      19.67  28.005  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 138.9 on 196 degrees of freedom
Multiple R-squared:  0.9358,        Adjusted R-squared:  0.9348
F-statistic: 951.7 on 3 and 196 DF,  p-value: < 2.2e-16
```

Modèle B :

```
> summary(lm(Sal~S+IndEtu+IndExp+S:IndEtu+S:IndExp))

Call:
lm(formula = Sal ~ S + IndEtu + IndExp + S:IndEtu + S:IndExp)

Residuals:
     Min       1Q   Median       3Q      Max
-225.057  -77.273   -2.623   72.424  261.169

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   819.70      27.73  29.562  < 2e-16 ***
S             155.63      38.53   4.039 7.73e-05 ***
IndEtu        108.22      36.53   2.963  0.00343 **
IndExp       1275.40      35.25  36.178  < 2e-16 ***
S:IndEtu      265.97      49.48   5.375 2.18e-07 ***
S:IndExp      530.59      51.25  10.354  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 107.1 on 194 degrees of freedom
Multiple R-squared:  0.9622,        Adjusted R-squared:  0.9612
F-statistic:   987 on 5 and 194 DF,  p-value: < 2.2e-16
```

FICHE T.D. 5 Prévision

Exercice 12 (suite des exercices 1 et 6)

Question 1 : Si en 2003, on comptabilisait 1800 films diffusés à la TV, pourriez-vous prévoir la fréquentation des salles de cinéma (en millions d'entrées) ? Déterminez un intervalle de prévision à 95% de votre prévision. Reportez, cette prévision ainsi que l'intervalle de prévision associé sur la figure de l'exercice 1. On pourra s'aider des sorties suivantes :

```
> x<-cbind(1,log(films))
> xTau<-c(1,log(1800))
```

Question 2 : Confrontez vos résultats avec la sortie `R` suivante :

```
> xTau<-data.frame(films=1800)
> predict.lm(lm(log(freq)~log(films)),xTau,interval="prediction")->yPred
> yPred
       fit      lwr      upr
1 4.790211 4.590833 4.989589
> exp(yPred)
       fit      lwr     upr
1 120.3267 98.57651 146.876
```

Annexe :

```
> freq
 [1] 328.3 311.7 291.2 275.8 259.1 234.7 211.4 203.2 183.9 184.4 177.0 184.4
[13] 176.0 179.4 181.7 177.3 170.3 178.5 178.1 174.8 189.2 201.9 198.8 190.8
[25] 175.0 167.8 136.7 124.7 120.8 121.7 117.5 116.0 132.7 124.4 130.2 136.7
[37] 149.0 170.1
> films
 [1]  117  107  106  224  285  275  342  337  349  382  392  370  460  444  459
[16]  517  526  524  537  527  500  475  475  485  500  950 1288 1330 1289 1362
[31] 1375 1398 1421 1434 1501 1513 1510 1554
```

FICHE T.D. 6 Examen d'économétrie (Janvier 2005)

Exercice 1

Lors d'un mémoire de DEA (promotion 2000-2001), Frédéric Rey et Alexandre Turpin ont étudié le taux de chômage. Ils ont recueilli un jeu de données (présenté à la fin du devoir) constitué de $n = 34$ observations annuelles (de 1960 à 1993). Voici une description des variables fournie par les auteurs du mémoire :

- `an` : année
- `chom` : taux de chômage.
- `txpib` : taux de variation du produit intérieur brut (pib) représentant le taux de croissance de l'économie.
- `deppub` : part des dépenses publiques par rapport au pib, qui peut ainsi représenter le degré d'intervention de l'état dans l'économie.
- `pfisc` : pressions fiscales, pour voir si une imposition trop importante des entreprises nuit à leur embauche et donc au niveau du chômage.
- `salva` : la part des salaires par rapport à la valeur ajoutée permettant de connaître l'influence du coût sur l'embauche.
- `infl` : taux d'inflation afin de vérifier la relation inverse entre le chômage et l'inflation définie par la courbe de Philips.

Partie I : modèle à un seul régresseur

On envisage un modèle linéaire expliquant la variable `chom` en fonction de la seule variable `txpib`. On tente une modélisation linéaire du type :

$$(chom)_i = \beta_0 + \beta_1 (txpib)_i + U_i, \qquad i = 1, \ldots, 34$$

Question 1 : Pourquoi les paramètres β_0 et β_1 ne sont pas calculables ?
Question 2 : On s'intéresse tout naturellement à l'estimation des paramètres β_0 et β_1. Déterminez les estimations obtenues par la méthode des moindres carrés.

On rappelle à titre indicatif que

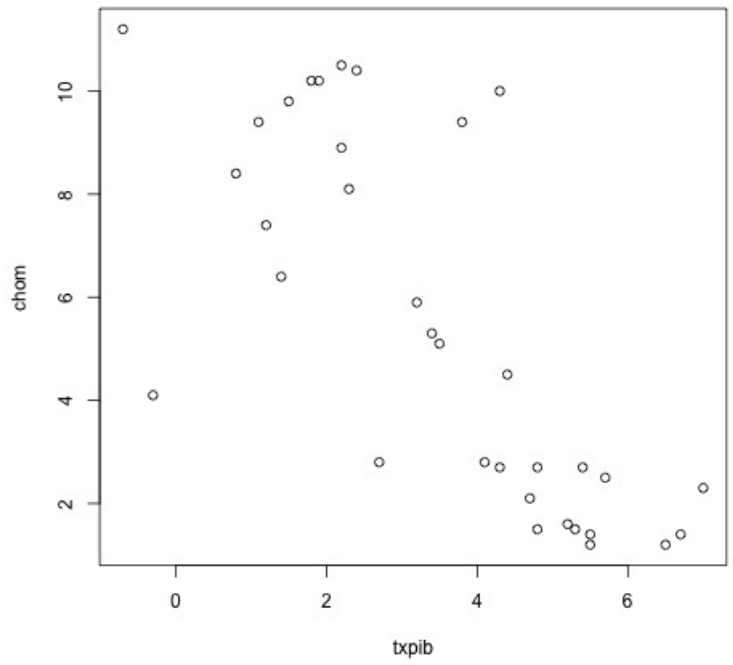
- $var(\boldsymbol{x}) = \overline{x^2} - \overline{x}^2$
- $cov(\boldsymbol{x}, \boldsymbol{y}) = \overline{x \times y} - \overline{x} \times \overline{y}$.

```
> mean(txpib)
[1] 3.488235
> mean(txpib^2)
[1] 16.06412
> mean(chom)
[1] 5.458824
> mean(chom^2)
[1] 42.03294
> mean(txpib*chom)
[1] 13.85794
```

Question 3 : Déterminez le coefficient de détermination linéaire (R^2) (puis le coefficient de corrélation linéaire (R)) entre `txpib` et `chom`, et donnez-en une interprétation.
Question 4 : Rappelez brièvement à quoi correspondent chacune des quatre colonnes de la matrice "Coefficients" de la sortie ci-dessous. Retrouvez les résultats des deux questions précédentes.

```
> summary(lm(chom~txpib))

Call:
lm(formula = chom ~ txpib)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3987 -1.5732 -0.2337  1.4000  5.6212

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.0996     0.8293   12.18 1.48e-13 ***
txpib        -1.3304     0.2069   -6.43 3.14e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.381 on 32 degrees of freedom
Multiple R-squared:  0.5637,        Adjusted R-squared:  0.5501
F-statistic: 41.34 on 1 and 32 DF,  p-value: 3.144e-07

> sqrt(0.5637051)
[1] 0.750803
```

Question 5 : Sur le graphique ci-dessous reportez la droite ajustée (même approximativement) et illustrez la notion de valeur ajustée et de résidu.



Question 6 : Peut-on penser au vu des données que la variable `txpib` apporte de l'information pour expliquer la variable `chom` (indication : fournir la p−valeur associée puis conclure.)

Partie II : modèle linéaire multiple

On envisage un modèle linéaire multiple expliquant la variable `chom` en fonction de tous les régresseurs du jeu de données (exceptée la variable `an`).

Question 1 : A la vue de la matrice de corrélation ci-après, quels sont les régresseurs qui vous semblent être les plus explicatifs ?

```
> cor(chomage[-1])
```

```
              chom      txpib       deppub       pfisc        salva        infl
chom    1.00000000 -0.7508029  0.978885951  0.97923991 -0.167503533 -0.04814577
txpib  -0.75080295  1.0000000 -0.780822861 -0.75374536 -0.105374429 -0.30407625
deppub  0.97888595 -0.7808229  1.000000000  0.99212248 -0.007785484  0.06750973
pfisc   0.97923991 -0.7537454  0.992122482  1.00000000 -0.035551565  0.06640145
salva  -0.16750353 -0.1053744 -0.007785484 -0.03555157  1.000000000  0.70322004
infl   -0.04814577 -0.3040763  0.067509734  0.06640145  0.703220044  1.00000000
```

Question 2 : Interprétez la sortie ci-dessous, en particulier les p−valeurs des tests de significativité locale, le R^2.

```
> summary(lm(chom~txpib+deppub+pfisc+salva+infl))

Call:
lm(formula = chom ~ txpib + deppub + pfisc + salva + infl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84262 -0.26630  0.05442  0.27937  1.33259

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.80283    3.86147  -1.503 0.144098
txpib       -0.07188    0.07753  -0.927 0.361820
deppub       0.35855    0.12702   2.823 0.008665 **
pfisc        0.22751    0.14324   1.588 0.123432
salva       -0.19195    0.05079  -3.779 0.000757 ***
infl        -0.03131    0.03961  -0.791 0.435873
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.469 on 28 degrees of freedom
Multiple R-squared:  0.9852,        Adjusted R-squared:  0.9826
F-statistic: 372.7 on 5 and 28 DF,  p-value: < 2.2e-16
```

Question 3 : Rappelez les effets indésirables sur les tests de significativité locale s'il y a colinéarité entre les régresseurs.
Question 4 : A la lumière de la matrice de corrélation associée au jeu de données, peut-on suspecter de la colinéarité entre les régresseurs ?
Question 5 : Rappelez la définition du VIF, et son interprétation générale. Ensuite, interprétez-les quant au jeu de données étudié.

```
> vif(lm(chom~txpib+deppub+pfisc+salva+infl))
    txpib    deppub     pfisc     salva      infl
 3.620834 91.775821 81.135288  2.384586  2.712915
```

Question 6 : (**Relation avec la matrice de corrélation**) Justifier l'ordre de grandeur des VIFs des covariables `deppub` et `pfisc` en utilisant l'instruction suivante.

```
> 1/(1-(0.992122482)^2)
[1] 63.72276
```

Question 7 : Quelle est la stratégie qui a été adoptée dans la série d'instructions ci-dessous ? A la dernière étape, précisez l'équation du modèle sélectionné et analysez brièvement les sorties.

```
> ## Etape 1
> summary(lm(chom~txpib+deppub+pfisc+salva))

Call:
lm(formula = chom ~ txpib + deppub + pfisc + salva)

```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.80961 -0.31312 -0.00256  0.26502  1.41147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.93872    3.03798  -1.296  0.20503
txpib       -0.04277    0.06779  -0.631  0.53302
deppub       0.39774    0.11619   3.423  0.00186 **
pfisc        0.18756    0.13315   1.409  0.16959
salva       -0.22177    0.03379  -6.563 3.44e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4659 on 29 degrees of freedom
Multiple R-squared:  0.9849,        Adjusted R-squared:  0.9828
F-statistic: 471.8 on 4 and 29 DF,  p-value: < 2.2e-16

> vif(lm(chom~txpib+deppub+pfisc+salva))
    txpib    deppub     pfisc     salva
 2.804297 77.798129 71.033914  1.069126
> ## Etape 2
> summary(lm(chom~deppub+pfisc+salva))

Call:
lm(formula = chom ~ deppub + pfisc + salva)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79521 -0.27194 -0.02836  0.26418  1.42664

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.63681    2.80081  -1.656 0.108244
deppub       0.42511    0.10670   3.984 0.000399 ***
pfisc        0.16761    0.12804   1.309 0.200456
salva       -0.21907    0.03318  -6.603 2.62e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4612 on 30 degrees of freedom
Multiple R-squared:  0.9847,        Adjusted R-squared:  0.9831
F-statistic: 641.8 on 3 and 30 DF,  p-value: < 2.2e-16

> vif(lm(chom~deppub+pfisc+salva))
   deppub     pfisc     salva
66.949894 67.030557  1.051973
> ## Etape 3
> summary(lm(chom~deppub+salva))

Call:
lm(formula = chom ~ deppub + salva)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73827 -0.36129 -0.03607  0.29649  1.37843

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -2.83284    2.46623  -1.149    0.259
deppub       0.56373    0.01319  42.741  < 2e-16 ***
salva       -0.22872    0.03272  -6.990 7.61e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4665 on 31 degrees of freedom
Multiple R-squared:  0.9838,        Adjusted R-squared:  0.9827
F-statistic: 940.2 on 2 and 31 DF,  p-value: < 2.2e-16

> vif(lm(chom~deppub+salva))
  deppub    salva
1.000061 1.000061

```

Question 8 : A partir de cette question, nous ne considèrerons que le modèle final. Peut-on montrer au vu des données que le paramètre $\beta_1 < 1$ au seuil de 5% ?

```
> (0.56373-(1))/0.01319
[1] -33.07582
```

Question 9 : A partir de l'instruction `R` suivante, que peut-on avancer au vu des données comme assertion(s) d'intérêt au seuil 5% ?

```
> pnorm((-0.22872-(-.3))/0.03272)
[1] 0.985315
```

Question 10 : En vous aidant de l'instruction `R` ci-dessous, proposez un intervalle de confiance à 95% pour le paramètre β_2 et interprétez-le (via l'approche expérimentale). Quelle relation y-a-t-il entre cet intervalle et le test de significativité locale du paramètre β_2 ?

```
> -0.22872+c(-1,1)*qnorm(0.975)*0.03272
[1] -0.29285 -0.16459
```

Question 11 : Supposons que l'on ne connaisse pas la valeur de `chom` en 1993. Pourriez-vous prévoir sa valeur, calculer un intervalle de prévision au niveau 95% ? Quelle était la valeur observée de `chom` en 1993 et est-ce surprenant ?

```
> xTau <- data.frame(chom=11.2,deppub=52.2,salva=68.6)
> predict(lm(chom~deppub+salva),xTau,interval="prediction")
       fit      lwr      upr
1 10.9039 9.872652 11.93515
```

Jeu de données :

```
> chomage
     an chom txpib deppub pfisc salva infl
1  1960  1.4   5.5   34.6  34.9  72.8  3.4
2  1961  1.2   5.5   35.7  36.2  72.8  3.4
3  1962  1.4   6.7   37.0  36.3  72.8  4.7
4  1963  1.5   5.3   37.8  37.1  72.8  6.4
5  1964  1.2   6.5   38.0  38.0  72.8  4.1
6  1965  1.5   4.8   38.4  38.4  73.3  2.7
7  1966  1.6   5.2   38.5  38.4  73.3  2.9
8  1967  2.1   4.7   39.0  38.2  73.3  3.2
9  1968  2.7   4.3   40.3  38.8  73.3  4.2
...
25 1984  9.8   1.5   52.5  49.8  75.6  7.3
26 1985 10.2   1.8   52.7  49.9  74.6  5.8
27 1986 10.4   2.4   52.2  49.4  72.1  5.3
28 1987 10.5   2.2   51.7  49.8  71.3  3.0
29 1988 10.0   4.3   50.8  49.2  70.2  3.1
```

```
30 1989  9.4   3.8   49.8  48.7  69.1  3.5
31 1990  8.9   2.2   50.3  48.9  69.6  3.1
32 1991  9.4   1.1   50.6  48.7  69.6  3.1
33 1992 10.2   1.9   51.3  48.5  68.8  2.9
34 1993 11.2  -0.7   52.2  49.0  68.6  2.9
```

FICHE T.D. 7 Examen d'économétrie (Janvier 2006)

Exercice 1

On envisage un modèle linéaire expliquant la variable `Chom` en fonction de la seule variable `Infl`. On tente une modélisation linéaire du type :

$$(Chom)_i = \beta_0 + \beta_1 (Infl)_i + U_i, \qquad i = 1, \ldots, 41$$

Question 1 : Pourquoi les paramètres β_0 et β_1 ne sont pas calculables ?
Question 2 : On s'intéresse tout naturellement à l'estimation des paramètres β_0 et β_1. Déterminez les estimations obtenues par la méthode des moindres carrés.

On rappelle à titre indicatif que

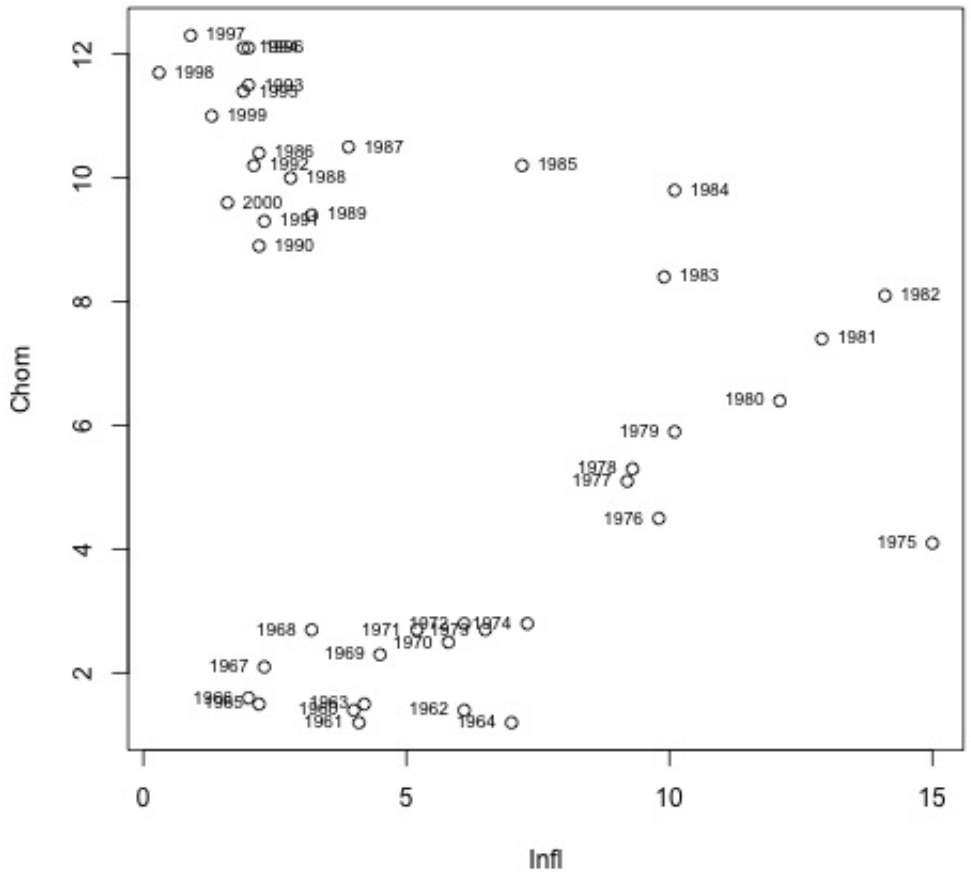
- $var(\boldsymbol{x}) = \overline{x^2} - \overline{x}^2$
- $cov(\boldsymbol{x}, \boldsymbol{y}) = \overline{x \times y} - \overline{x} \times \overline{y}$.

```
> mean(Infl)
[1] 5.385366
> mean(Infl^2)
[1] 44.03463
> mean(Chom)
[1] 6.487805
> mean(Chom^2)
[1] 57.51659
> mean(Infl*Chom)
[1] 31.02854
```

Question 3 : Déterminez le coefficient de corrélation linéaire entre `Infl` et `Chom` et donnez-en une interprétation.
Question 4 : Retrouvez les résultats des deux questions précédentes. Analysez brièvement les résultats obtenus.

```
> summary(lm(Chom~Infl))

Call:
lm(formula = Chom ~ Infl)

Residuals:
   Min     1Q Median     3Q    Max
-5.816 -3.880  1.583  3.449  4.731

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8888     1.0403   7.583 3.43e-09 ***
Infl         -0.2601     0.1568  -1.659    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.892 on 39 degrees of freedom
Multiple R-squared:  0.06595,        Adjusted R-squared:  0.042
F-statistic: 2.754 on 1 and 39 DF,  p-value: 0.105

> sqrt(0.06595486)
[1] 0.2568168
```

Question 5 : Sur le graphique ci-dessous reportez la droite ajustée (même approximativement) et illustrez la notion de valeur ajustée et de résidu.



Question 6 : Peut-on penser au vu des données que la variable `Infl` apporte de l'information pour expliquer la variable `Chom` (indication : fournir la p−valeur associée puis conclure.)
Question 7 : Quelle variable nommée `z` (ayant pour modalités P1 et P2) dans la sortie suivante a été introduite ? Comparez les résultats obtenus avec ceux du précédent modèle. Représentez sur le graphique précédent ce nouveau modèle dont on précisera l'équation. Peut-on alors penser que `Infl` a un pouvoir explicatif sur `Chom` ?

```
> summary(lm(Chom~Infl*z))

Call:
lm(formula = Chom ~ Infl * z)

Residuals:
     Min       1Q   Median       3Q      Max
-1.99808 -0.59426  0.09453  0.71012  1.61847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.53721    0.48284   1.113    0.273
Infl         0.37072    0.06625   5.595 2.22e-06 ***
zP2         10.71149    0.58361  18.354  < 2e-16 ***
Infl:zP2    -0.61452    0.08663  -7.094 2.13e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```

```
Residual standard error: 1.015 on 37 degrees of freedom
Multiple R-squared:  0.9398,        Adjusted R-squared:  0.9349
F-statistic: 192.4 on 3 and 37 DF,  p-value: < 2.2e-16
```

Exercice 2

L'étude portera sur l'analyse du prix d'une maison en fonction de ses différentes caractéristiques, et sera basée sur le jeu de données `maison` de taille $n = 150$ présenté à la fin du devoir. Voici très brièvement la description des variables associées :

- `PRIX` : prix des maisons.
- `SURFACE` : suface des maisons.
- `HECTARES` : surfaces des terrains associés.
- `PIECES` : nombre de pièces des maisons.
- `BAINS` : nombre de salle de bains.

On envisage un modèle log-linéaire multiple expliquant la variable `log(PRIX)` en fonction de tous les régresseurs du jeu de données.

Question 1 : A la vue de la matrice de corrélation ci-après, quels sont les régresseurs qui vous semblent être les plus explicatifs ?

```
> cor(log(maison))
              PRIX   SURFACE  HECTARES    PIECES     BAINS
PRIX     1.0000000 0.7746669 0.5374595 0.6612338 0.7089474
SURFACE  0.7746669 1.0000000 0.3118338 0.8542915 0.7856913
HECTARES 0.5374595 0.3118338 1.0000000 0.1961679 0.3525309
PIECES   0.6612338 0.8542915 0.1961679 1.0000000 0.6502278
BAINS    0.7089474 0.7856913 0.3525309 0.6502278 1.0000000
```

Question 2 : Interprétez la sortie ci-dessous, en particulier les p−valeurs des tests de significativité locale, le R^2.

```
> summary(lm(log(PRIX)~log(SURFACE)+log(HECTARES)+log(PIECES)+log(BAINS)))

Call:
lm(formula = log(PRIX) ~ log(SURFACE) + log(HECTARES) + log(PIECES) +
    log(BAINS))

Residuals:
     Min       1Q   Median       3Q      Max
-0.42859 -0.08990  0.00517  0.08221  0.36644

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.00815    0.47661  18.901  < 2e-16 ***
log(SURFACE)   0.35710    0.08354   4.275 3.45e-05 ***
log(HECTARES)  0.09231    0.01419   6.504 1.19e-09 ***
log(PIECES)    0.11086    0.10312   1.075    0.284
log(BAINS)     0.11738    0.04817   2.437    0.016 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1419 on 145 degrees of freedom
Multiple R-squared:  0.7108,        Adjusted R-squared:  0.7029
F-statistic: 89.11 on 4 and 145 DF,  p-value: < 2.2e-16
```

Question 3 : Rappelez les effets indésirables sur les tests de significativité locale s'il y a colinéarité entre les régresseurs.

Question 4 : Rappelez la définition du VIF, et son interprétation générale. Ensuite, interprétez-les quant au jeu de données étudié.

```
> vif(lm(log(PRIX)~log(SURFACE)+log(HECTARES)+log(PIECES)+log(BAINS)))
 log(SURFACE) log(HECTARES)   log(PIECES)    log(BAINS)
     5.715257      1.166638      3.783690      2.707968
```

Question 5 : (**Relation avec la matrice de corrélation**) Justifier l'ordre de grandeur des VIFs des covariables `log(SURFACE)` et `log(PIECES)` en utilisant l'instruction suivante.

```
> 1/(1-(0.854291515)^2)
[1] 3.701154
```

Question 6 : Quelle est la stratégie à adopter pour soigner la colinéarité ? En particulier, rappelez son effet sur les erreurs standard. Après une seule étape voici les résultats du `summary(lm(...))` présentés sous une forme spécialement adaptée aux notations du cours. Complétez les 3 **cases manquantes** en vous aidant des indications (à la suite du tableau).Justifiez qu'il n'est pas nécessaire d'effectuer d'étape supplémentaire et précisez l'équation du modèle sélectionné.

$\mathbf{x}^{(j)}$	$\widehat{\beta_j}(\mathbf{y}\|\underline{\mathbf{x}})$	$\widehat{\sigma}_{\widehat{\beta_j}}(\mathbf{y}\|\underline{\mathbf{x}})$	$\widehat{\delta_{\beta_j,0}}(\mathbf{y}\|\underline{\mathbf{x}})$	p-valeur bilatérale
1	8.73297	0.40225	21.71048	$\simeq 0$
log(SURFACE)	0.42316			$\simeq 0$
log(HECTARES)	0.09028	0.01408	6.41437	$\simeq 0$
	0.11529	0.04816	2.39405	0.01793

Indications : $R^2 = 70.85000000000001\%$, $\widehat{\sigma}_U(\mathbf{y}|\underline{\mathbf{x}}) \simeq 0.142$ et $VIF_{PRIX} \simeq 2.6226$.

Question 7 : Etant donnée la modélisation adoptée, complétez la phrase ci-dessous : lorsque `SURFACE` _____ de 10% on peut s'attendre à ce que `log(PRIX)` _____ de ___.

Question 8 : A partir de cette question, nous ne considèrerons que le modèle final. Peut-on montrer au vu des données que le paramètre $\beta_1 < 0.5$ au seuil de 5% ?

```
> (0.42316-(0.5))/0.05662
[1] -1.357118
```

Question 9 : A partir de l'instruction `R` suivante, que peut-on avancer au vu des données comme assertion(s) d'intérêt au seuil 5% ?

```
> pnorm((0.09028-(0.05))/0.01408)
[1] 0.9978871
```

Question 10 : En vous aidant de l'instruction `R` ci-dessous, proposez un intervalle de confiance à 95% pour le paramètre β_2 et interprétez-le (via l'approche expérimentale). Quelle relation y-a-t-il entre cet intervalle et le test de significativité locale du paramètre β_2 ?

```
> 0.09028+c(-1,1)*qnorm(0.975)*0.01408
[1] 0.06268371 0.11787629
```

Jeu de données :

```
> maison
      PRIX SURFACE HECTARES PIECES BAINS
1   179000    3060   0.7500      8   2.0
2   126500    1600   0.2600      8   1.5
3   134500    2000   0.7000      8   1.0
4   125000    1300   0.6500      5   1.0
5   142000    2000   0.7500      9   1.5
6   164000    1956   0.5000      8   2.5
7   146000    2400   0.4000      7   2.5
8   129000    1200   0.3300      6   1.0
9   141900    1632   3.0000      6   3.0
...
141 121900    1300   0.7800      6   1.0
```

```
142 126000    1232   0.3140      6   2.0
143 164900    1980   0.7000      8   2.5
144 172000    2100   1.0000      8   2.5
145 100000    1338   0.1200      6   1.0
146 129900    1070   1.6900      5   1.0
147 110000    1289   0.2500      6   1.0
148 131000    1066   0.3300      5   1.0
149 107000    1100   0.1700      5   1.0
150 165900    1840   1.1620      8   2.0
```