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 :

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

Question 1: *Pourquoi les paramètres β_0 et β_1 ne sont pas calculables ?*

Réponse

Pour pouvoir évaluer les paramètres β_0 et β_1 il faudrait "en théorie" une infinité de données. Ces paramètres ne sont donc pas évaluables mais simplement estimables à partir d'un nombre fini de données.

Fin

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
```

Réponse

En notant $\boldsymbol{y} = \boldsymbol{Chom}$ et $\underline{\boldsymbol{x}} = (\mathbf{1}, \boldsymbol{x^{(1)}} := \boldsymbol{Infl})$,

$$\widehat{\beta_1}\,(\boldsymbol{y}|\underline{\boldsymbol{x}}) = \frac{cov(\boldsymbol{y}, \boldsymbol{x^{(1)}})}{var(\boldsymbol{x^{(1)}})} = \frac{31.02854 - (5.385366) * (6.487805)}{44.03463 - 5.385366^2} \simeq -0.2601480$$

$$\widehat{\beta_0}\,(\boldsymbol{y}|\underline{\boldsymbol{x}}) = \overline{\boldsymbol{y}} - \widehat{\beta_1}\,(\boldsymbol{y}|\underline{\boldsymbol{x}})\,\overline{\boldsymbol{x^{(1)}}} = 6.487805 - -0.2601480 \times 5.385366 \simeq 7.888797.$$

Vérification en `R` :

```
> beta1Est<-(mean(Infl*Chom)-mean(Infl)*mean(Chom))/(mean(Infl^2)-mean(Infl)^2)
> beta1Est
[1] -0.2601480
> mean(Chom) - beta1Est*mean(Infl)
[1] 7.888797
```

Fin

Question 3: *Déterminez le coefficient de corrélation linéaire entre `Infl` et `Chom` et donnez-en une interprétation.*

Réponse

Le coefficient de corrélation linéaire vaut

$$corr(\boldsymbol{y}, \boldsymbol{x}) = \frac{cov(\boldsymbol{y}, \boldsymbol{x^{(1)}})}{\sqrt{var(\boldsymbol{y}) \times var(\boldsymbol{x^{(1)}})}} = \frac{31.02854 - (5.385366) * (6.487805)}{\sqrt{(44.03463 - 5.385366^2)(57.51659 - 6.487805^2)}} \simeq -0.2568168,$$

et plus il est proche de 1 en valeur absolue plus la dispersion des points autour de la droite ajustée est faible.

Fin

Question 4: *Retrouvez les résultats des deux questions précédentes. Analysez brièvement les résultats obtenus.*

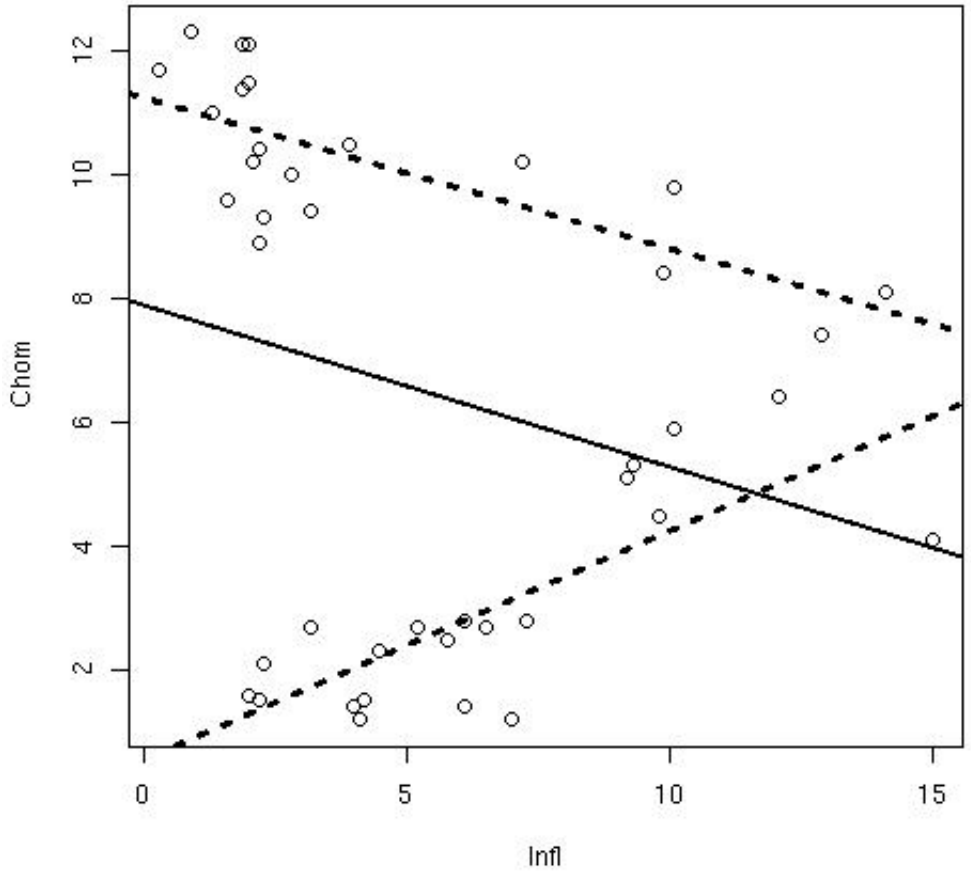
```
> summary(lm(Chom~Infl))

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: 0.1050

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

Réponse

On retrouve les résultats des deux questions précédentes. En particulier, la colonne `Estimate` fournit les estimations des paramètres β_0 et β_1 et `Multiple R-squared` le R^2. Notons que la dernière colonne du tableau `Coefficients` fournit les p−valeurs des tests de significativité locales des paramètres β_0 et β_1. La dernière ne laisse pas en particulier apparaître que la variable `Infl` semble apporter de l'information dans l'explication de la variable `Chom`.

Fin

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



Les notions de résidu et de valeur ajustée sont illustrés dans le polycopié p.23.

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.)*

Réponse

Pour pouvoir accepter $\mathbf{H_1}$: $\beta_1 \neq 0$, i.e. le régresseur `Infl` apporte de l'information pour expliquer la variable `Chom`, le risque à encourir est de l'ordre de 0.1050409.

Fin

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
```

Réponse

La variable nommée z correspond très certainement avec la variable indicatrice définie par : $z = P1$ si $An < 1981$ et $z = P2$ sinon. L'équation de ce modèle s'écrit

$$(Chom)_i = \beta_0 + \beta_1(Infl)_i + \beta_2(z)_i + \beta_3(Infl * z)_i + \varepsilon_i.$$

Le modèle estimé correspond à deux droites d'équation $y = (\widehat{\beta_0} + \widehat{\beta_2}) + (\widehat{\beta_1} + \widehat{\beta_3})x$ lorsque $z = 1$ et $y = \widehat{\beta_0} + \widehat{\beta_1}x$ lorsque $X = 0$ (ces deux droites sont représentées en pointillé sur le graphique précédent). On notera que pour ce nouveau modèle, le pouvoir explicatif exprimé par le R^2 est de l'ordre de 93.98 soit une augmentation de près de 87.38% par rapport au précédent modèle. Ainsi, par le simple ajout de la variable indicatrice `z`, on s'aperçoit que la variable possède un fort pouvoir explicatif de la variable `Chom`.

Fin

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
```

Réponse

A la vue de la matrice de corrélation, les régresseurs les plus explicatifs de la variabilité de la variable `log(PRIX)` semblent être dans l'ordre `SURFACE`, `BAINS`, `PIECES` et `HECTARES` (en ne tenant compte que des régresseurs ayant un coefficient de corrélation avec `log(PRIX)` en valeur absolue supérieur à 30.0%).

Fin

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   0.2842
log(BAINS)     0.11738    0.04817   2.437   0.0160 *
---
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
```

Réponse

Les régresseurs `log(HECTARES)`, `log(SURFACE)` et `log(BAINS)` sont significatifs au seuil de 5%. Enfin, le pouvoir explicatif de ce modèle, mesuré par la part de variance R^2 expliquée par celui-ci, est particulièrement bon puisqu'il est de l'ordre de 71.08383%.

Fin

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

Réponse

La colinéarité entre régresseurs peut entraîner artificiellement une augmentation de l'erreur standard des paramètres des régresseurs et par voie de fait une diminution en valeur absolue de la statistique de test de significativité locale et donc une augmentation de la p−valeur de ce même test.

Fin

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
```

Réponse

Par définition, $VIF_j = \frac{1}{1-R_j^2}$ où R_j^2 est le coefficient de détermination multiple au carré lorsque l'on régresse le $j-$ème régresseur sur tous les autres. Par ailleurs, on sait d'après le cours que l'erreur standard de l'estimateur de β_j est proportionnelle au VIF_j. Par conséquent, plus le $j-$ème régresseur est corrélé avec les autres, plus R_j^2 est proche de 1, plus l'erreur standard de l'estimateur de β_j sera grande et plus il sera difficile de montrer que ce régresseur apporte de l'information. Le VIF de `log(SURFACE)` est relativement important, ici supérieur à 5. Quand il existe un VIF relativement important, cela traduit une forte colinéarité entre régresseurs.

Fin

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
```

Réponse

Il est connu que lorsque l'on ajoute des régresseurs le coefficient R^2 augmente nécessairement. Ainsi par exemple, le R^2 obtenu en régressant `SURFACE` sur tous les autres ou `PIECES` sur tous les autres est nécessairement supérieurs au R^2 obtenu en régressant simplement `SURFACE` sur `PIECES` (de l'ordre de 0.854291515^2). Par conséquent les VIF associées à ces deux régresseurs sont nécessairement supérieurs à 3.701154.

Fin

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	**0.05662**	**7.47405**	$\simeq 0$
log(HECTARES)	0.09028	0.01408	6.41437	$\simeq 0$
log(BAINS)	0.11529	0.04816	2.39405	0.01793

Indications : $R^2 = 70.85\%$, $\widehat{\sigma}_{\varepsilon}(\mathbf{y}|\underline{\mathbf{x}}) \simeq 0.142$ et $VIF_{PRIX} \simeq 2.6226$.

Réponse

Fin

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
```

Réponse

D'après ce qui précède, la statistique du test $\mathbf{H_1} : \beta_1 < 0.5$ évaluée sur les données correspond justement au calcul fourni et vaut donc approximativement -1.357118. Sous $\mathbf{H_0}$ on sait également que la statistique de test sur les futures données suit approximativement une loi $\mathcal{N}(0,1)$. Par conséquent, puisque $\widehat{\delta_{\beta_1,0.5}}(\boldsymbol{y}|\underline{\boldsymbol{x}}) \simeq -1.357118 > \delta^-_{lim,5\%} \overset{R}{=} \texttt{qnorm}(.05) \simeq -1.644854$, on ne peut pas plutôt penser avec un risque de 5% que l'assertion d'intérêt $\mathbf{H_1} : \beta_1 < 0.5$ est vraie.

Fin

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
```

Réponse

L'instruction fournie correspond à la p−valeur (gauche) du test $\mathbf{H_1} : \beta_2 < 0.05$. Par conséquent puisque la p−valeur droite associée au test $\mathbf{H_1} : \beta_2 > 0.05$ qui vaut approximativement 0.2104988% est inférieure à 5% , on peut plutôt penser que cette assertion est vraie. Notons, qu'il est possible de penser que $\mathbf{H_1} : \beta_2 \neq 0.05$ est vraie car la p−valeur de ce test vaut approximativement $2 \times 0.2104988\% = 0.4209977\%$.

Fin

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
```

Réponse

Fin

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
```