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 :

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

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(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
```

Réponse

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

$$\widehat{\beta_1}\left(\boldsymbol{y}|\underline{\boldsymbol{x}}\right) = \frac{cov(\boldsymbol{y},\boldsymbol{x^{(1)}})}{var(\boldsymbol{x^{(1)}})} = \frac{13.85794-(3.488235)*(5.458824)}{16.06412-3.488235^2} \simeq -1.33041$$
$$\widehat{\beta_0}\left(\boldsymbol{y}|\underline{\boldsymbol{x}}\right) = \overline{\boldsymbol{y}} - \widehat{\beta_1}\left(\boldsymbol{y}|\underline{\boldsymbol{x}}\right)\overline{\boldsymbol{x^{(1)}}} = 5.458824 - -1.33041 \times 3.488235 \simeq 10.09961.$$

Vérification en `R` :

```
> beta1Est<-(mean(txpib*chom)-mean(txpib)*mean(chom))/
+ (mean(txpib^2)-mean(txpib)^2)
> beta1Est
[1] -1.33041
> mean(chom) - beta1Est*mean(txpib)
[1] 10.09961
```

Fin

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.*

Réponse

Dans le cadre de la régression simple, R^2 vaut le carré du coefficient de corrélation linéaire :

$$R^2 = \frac{cov(\boldsymbol{y},\boldsymbol{x^{(1)}})^2}{var(\boldsymbol{y}) \times var(\boldsymbol{x^{(1)}})} = \frac{(13.85794-(3.488235)*(5.458824))^2}{(16.06412-3.488235^2)(42.03294-5.458824^2)} \simeq 0.5637051.$$

D'où $corr(\boldsymbol{x},\boldsymbol{y}) = -\sqrt{0.5637051} \simeq -0.750803$ qui a en particulier le même signe que $\widehat{\beta_1}\left(\boldsymbol{y}|\underline{\boldsymbol{x}}\right)$. Plus $|corr(\boldsymbol{x},\boldsymbol{y})|$ est proche de 1 et plus la dispersion des points autour de la droite ajustée est faible.

Fin

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
```

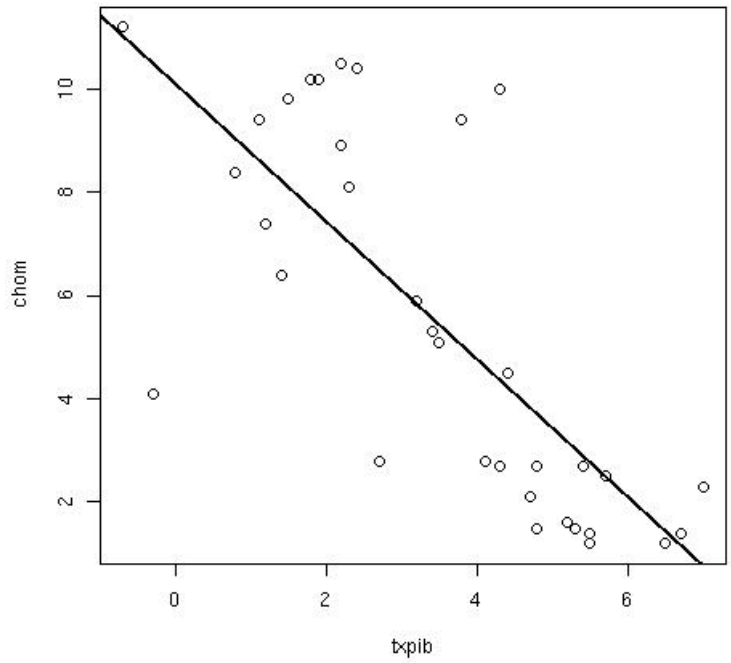
Réponse

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 laisse en particulier apparaître que la variable `txpib` 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* `txpib` *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 `txpib` apporte de l'information pour expliquer la variable `chom`, le risque à encourir est de l'ordre de $3.144288e-07$.

Fin

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
```

Réponse

A la vue de la matrice de corrélation, les régresseurs les plus explicatifs de la variabilité de la variable `chom` semblent être dans l'ordre `pfisc`, `deppub` et `txpib` (en ne tenant compte que des

régresseurs ayant un coefficient de corrélation avec `chom` 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(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
```

Réponse

Les régresseurs `salva` et `deppub` 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 98.51961%.

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: *A la lumière de la matrice de corrélation associée au jeu de données, peut-on suspecter de la colinéarité entre les régresseurs ?*

Réponse

Fin

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
```

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. Les VIFs de `pfisc` et `deppub` sont relativement importants, ici supérieurs à 5. Quand il existe un VIF relativement important, cela traduit une forte colinéarité entre régresseurs.

Fin

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
```

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 `deppub` sur tous les autres ou `pfisc` sur tous les autres est nécessairement supérieurs au R^2 obtenu en régressant simplement `deppub` sur `pfisc` (de l'ordre de 0.992122482^2). Par conséquent les VIF associées à ces deux régresseurs sont nécessairement supérieurs à 63.72276.

Fin

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  7.6e-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
```

Réponse

Fin

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
```

Réponse

D'après ce qui précède, la statistique du test $\mathbf{H_1} : \beta_1 < 1$ évaluée sur les données correspond justement au calcul fourni et vaut donc approximativement -33.07582. 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 $\widehat{\delta_{\beta_1,1}}(\boldsymbol{y}|\underline{\boldsymbol{x}}) \simeq -33.07582 < \delta^{-}_{lim,5\%} \overset{R}{=} \texttt{qnorm}(.05) \simeq -1.644854$, on peut plutôt penser avec un risque de 5% que l'assertion d'intérêt $\mathbf{H_1} : \beta_1 < 1$ 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((-0.22872-(-.3))/0.03272)
[1] 0.985315
```

Réponse

L'instruction fournie correspond à la p−valeur (gauche) du test $\mathbf{H_1} : \beta_2 < -.3$. Par conséquent puisque la p−valeur droite associée au test $\mathbf{H_1} : \beta_2 > -.3$ qui vaut approximativement 1.468376% est inférieure à 5% , on peut plutôt penser que cette assertion est vraie. Notons, qu'il est possible de penser que $\mathbf{H_1} : \beta_2 \neq -.3$ est vraie car la p−valeur de ce test vaut approximativement $2 \times 1.468376\% = 2.936751\%$.

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 ?*

```
> -0.22872+c(-1,1)*qnorm(0.975)*0.03272
[1] -0.2928500 -0.1645900
```

Réponse

Fin

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
```

Réponse

Fin

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
```