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 :

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

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

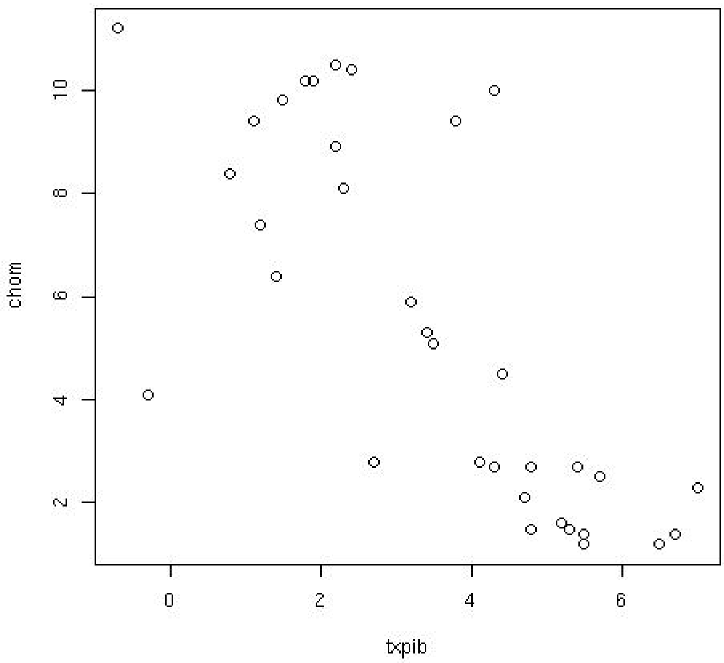
- $var(\boldsymbol{x}) = \overline{x^2} - \overline{x}^2$
- $cov(\boldsymbol{x}, \boldsymbol{y}) = \overline{x \times y} - \overline{x} \times \overline{y}$.

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



Question : *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 : *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 : *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 : *Rappelez les effets indésirables sur les tests de significativité locale s'il y a colinéarité entre les régresseurs.*
Question : *A la lumière de la matrice de corrélation associée au jeu de données, peut-on suspecter de la colinéarité entre les régresseurs ?*
Question : *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 : *(**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 : *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
-0.809610 -0.313125 -0.002559  0.265016  1.411467

```

```
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 

```

Question : *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 : *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 : *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
```

Question : *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
```