Université Paris Nanterre
Statistiques L3 Gestion Apprentissage

Exercices :
Modèle de régression linéaire simple et multiple

Exercice 1 *On a relevé pour différents pays le PIB par habitant en 2004 X (en dollars) et le taux brut de scolarisation des moins de 24 ans la même année Y (en pourcentage). Les résultats sont les suivants*

Pays	*PIB X*	*Taux de scolarisation Y*
Pays en développement	*4775*	*63*
Pays les plus pauvres	*1350*	*45*
Pays arabes	*5680*	*62*
Asie de l'Est et Pacifique	*5872*	*69*
Amérique latine et Caraïbes	*7964*	*81*
Asie du Sud	*3072*	*56*
Afrique Sub-saharienne	*1942*	*50*
Europe centrale,orientale et CEI	*8802*	*83*

$$\sum x_i = 39457; \quad \sum y_i = 509;$$
$$\sum x_i^2 = 245474957; \quad \sum y_i^2 = 33685; \quad \sum x_i y_i = 2763685$$

1. *On cherche à expliquer le taux de scolarisation en fonction du PIB. Identifier la variable à expliquer et la variable variable explicative. Pour chaque variable calculer la moyenne observée et la variance observée.*

2. *Expliquer l'objectif de la régression linéaire simple et préciser ses conditions d'application. Donner l'équation du modèle théorique.*

3. *Donner l'équation de la droite avec les valeurs estimées des coefficients inconnus β_0 et β_1.*

Indications :

$$\bar{x} = \frac{\sum x_i}{n} \quad \bar{y} = \frac{\sum y_i}{n} \quad \mathrm{cov}(x,y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n-1} \quad \mathrm{s}_x = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} \quad \mathrm{s}_y = \sqrt{\frac{\sum y_i^2 - n\bar{y}^2}{n-1}}$$

$$\mathrm{r}(x,y) = \frac{\mathrm{cov}}{\mathrm{s}_x \mathrm{s}_y} \quad \hat{\beta}_1 = \frac{\mathrm{r}(x,y) * \mathrm{s}_y}{\mathrm{s}_x} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}.$$

Exercice 2 *À environnement (quartier ou ville) donné, une idée généralement partagée est que la surface d'un appartement détermine assez largement son prix. Sans aucun doute, la surface d'un appartement et son prix sont très fortement liés. Nous souhaitons donc expliquer le prix en kilo euros en fonction de la surface en m^2.*

Nous disposons d'un échantillon $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ de taille $n = 28$ où x_i représente la surface de l'appartement i et y_i son prix.

Pour modéliser la dépendance entre le prix d'un appartement et la surface, nous choisissons le modèle de la régression linéaire simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \textit{pour tout} \quad i = 1, \ldots, n.$$

1. *Que représentent respectivement les termes $\beta_0 + \beta_1 x_i$ et ε_i dans l'équation ci-dessous ?*

2. *Quelle est la méthode qui permet d'estimer les coefficients β_0 et β_1 ? Expliquer très brièvement le principe de cette méthode (motiver également ce problème à l'aide d'un graphique).*

3. *Nous avons ajusté un modèle de régression linéaire simple pour expliquer le prix en fonction de la surface.*

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -29.466     41.245  -0.714    0.481
surface        5.353      0.414  12.931 7.86e-13
---

Residual standard error: 122.9 on 26 degrees of freedom
Multiple R-squared:  0.8654,Adjusted R-squared:  0.8603
F-statistic: 167.2 on 1 and 26 DF,  p-value: 7.862e-13
```

   (a) *Quelle est la variable à expliquer? Quelle est la variable explicative ?*

   (b) *Donner les estimations des coefficients de la régression et préciser leur interprétation.*

   (c) *Donner l'équation de la droite ajustée.*

   (d) *Tester la nullité de la pente de la droite de régression en précisant les hypothèses nulle et alternative du test. Que conclure au seuil 5%?*

   (e) *Relever la valeur observée du coefficient de détermination R^2 et l'interpréter.*

4. *Expliquer comment on obtient les deux derniers lignes du tableau ci-dessous (prix prédit et Résidus).*

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
prix observée	*130.00*	*280.00*	*650.00*	*800.00*	*268.00*	*790.00*	*500.00*	*320.00*
prix prédit	*120.42*	*238.19*	*537.97*	*1019.75*	*264.96*	*987.64*	*559.38*	*291.72*
Résidus	*9.58*	*41.81*	*112.03*	*-219.75*	*3.04*	*-197.64*	*-59.38*	*28.28*

Exercice 3 *Cet exercice porte sur les données observées sur un échantillon de 474 employés tirés au sort dans une entreprise canadienne. Les variables étudiées ici sont les suivantes :*

- `salary` *(salaire brut actuel en $ par an)*
- `salbegin` *(salaire de départ en $ par an)*
- `jobtime`*(nombre de mois depuis l'entrée dans l'entreprise)*
- `prevexp` *(nombre de mois de travail avant l'entrée dans l'entreprise)*
- `educ` *(nombre d'années d'étude)*
- `sex` *(sexe à deux modalités H = Homme et F = Femme)*

On souhaite expliquer la variable `salary` *en fonction de toutes les autres variables (*`salbegin`*,* `jobtime`*,* `prevexp`*,* `educ` *et* `sex`*) à l'aide de la régression linéaire.*

1. *Nous avons déterminé la matrice de corrélation.*

```
salary    salbegin      jobtime      prevexp        educ
salary    1.00000000  0.88011747  0.084092267 -0.097466926  0.66055891
salbegin  0.88011747  1.00000000 -0.019753475  0.045135627  0.63319565
jobtime   0.08409227 -0.01975347  1.000000000  0.002978134  0.04737878
prevexp  -0.09746693  0.04513563  0.002978134  1.000000000 -0.25235252
educ      0.66055891  0.63319565  0.047378777 -0.252352521  1.00000000
```

   (a) *Indiquer pour quels couples de variables la corrélation linéaire observée est la plus forte, la plus faible. Que peut-on dire de la corrélation linéaire entre le salaire de départ et le salaire actuel ?*

   (b) *Pourquoi n'y a-t-il pas la variable* ***sex*** *dans le matrice de corrélation ?*

2. *Nous avons ajusté un modèle de régression linéaire multiple expliquant* ***salary*** *en fonction de toutes les autres (****salbegin****,* ***jobtime****,* ***prevexp****,* ***educ*** *et* ***sex****).*

   Modèle 1 :

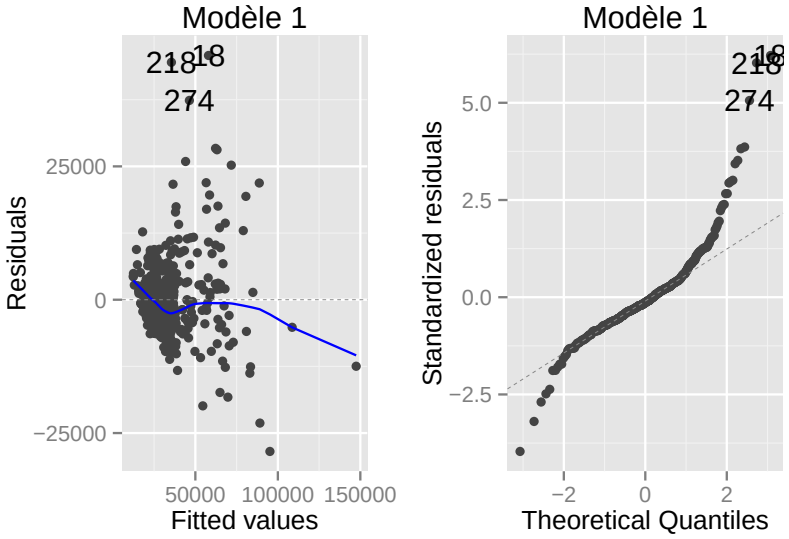
```
Modele1=lm(formula = salary ~ salbegin + jobtime + prevexp + educ + sex,
data = Salaire)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.255e+04  3.475e+03  -3.612 0.000337 ***
salbegin     1.723e+00  6.051e-02  28.472  < 2e-16 ***
jobtime      1.545e+02  3.408e+01   4.534 7.37e-06 ***
prevexp     -1.944e+01  3.583e+00  -5.424 9.36e-08 ***
educ         5.930e+02  1.666e+02   3.559 0.000410 ***
sexF        -2.233e+03  7.921e+02  -2.819 0.005021 **
---
Residual standard error: 7410 on 468 degrees of freedom
Multiple R-squared:  0.8137,Adjusted R-squared:  0.8117
F-statistic: 408.7 on 5 and 468 DF,  p-value: < 2.2e-16
```

   (a) *Tester la significativité globale du modèle à un niveau de risque de* 5% *en n'oubliant pas de donner les hypothèses nulles et alternatives du test. Que peut-on conclure ?*

   (b) *Relever et interpréter la valeur observée du coefficient* R^2*.*

   (c) *Quelles sont les variables significatives au seuil de signification de* 5% *?*

   (d) *Que représentent les graphes ci-dessous ?*



(e) Pensez vous que le modèle ajusté est pertinent ? Justifier.

3. *Nous avons appliqué une transformation logarithmique aux variables* ***salary*** *et* ***salbegin*** *et nous avons ajusté un modèle de régression linéaire multiple en remplaçant ces variables par les variables transformées.*

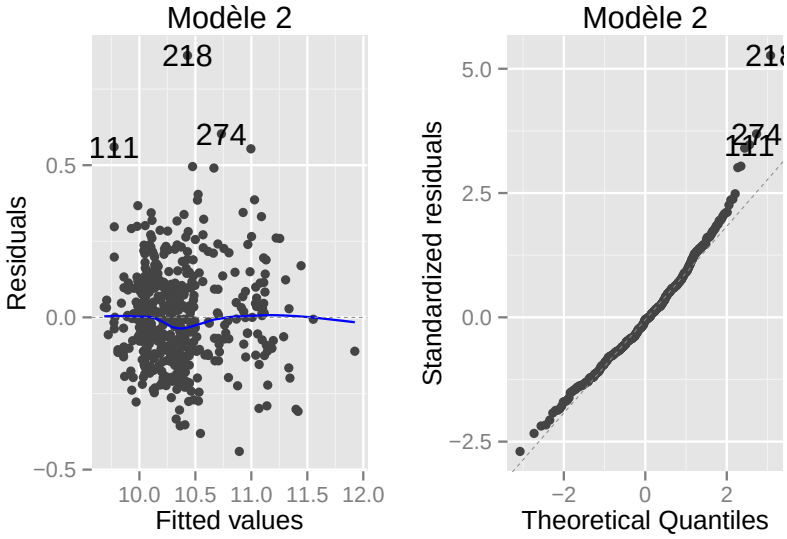
Modèle 2 :

```
Modele2=lm(formula = log(salary) ~ log(salbegin) + jobtime + prevexp +
educ + sex, data = Salaire)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.116e+00  3.125e-01   3.571 0.000392 ***
log(salbegin)  9.107e-01  3.382e-02  26.924  < 2e-16 ***
jobtime        4.517e-03  7.579e-04   5.960 4.97e-09 ***
prevexp       -5.527e-04  7.932e-05  -6.968 1.10e-11 ***
educ           1.071e-02  3.912e-03   2.737 0.006431 **
sexF          -4.995e-02  1.844e-02  -2.708 0.007019 **
---
Residual standard error: 0.1639 on 468 degrees of freedom
Multiple R-squared:  0.8317,Adjusted R-squared:  0.8299
F-statistic: 462.6 on 5 and 468 DF,  p-value: < 2.2e-16
```

(a) Tester la significativité globale du modèle à un niveau de risque de 5% en n'oubliant pas de donner les hypothèses nulles et alternatives du test. Que peut-on conclure ?

(b) Relever et interpréter la valeur observée du coefficient R^2.

(c) Peut-on valider ce modèle ? Expliquer.



4. *Lequel des deux modèles de régression multiple considérés préférez vous ? Appuyez votre réponse sur les graphes pertinents.*

5. *Donner l'équation du modèle ajusté que vous avez choisi.*

Exercice 4 *Cet exercice porte sur les données de la mortalité routière en Europe. Nous disposons d'un échantillon de taille 27. Les variables étudiées ici sont les suivantes :*

- `MortsPM`*: Mortalité sur les routes par million selon les données de l'UE.*
- `Transp`*: Transparence selon Heritage Foundation.*
- `Alcool`*: Taux d'alcoolémie permis par la loi.*
- `Nvdemo`*: Nouvelle démocratie, Ancienne démocratie.*

Il est important remarquer que le terme Transparence utilisé dans la variable `Transp` *correspond à un indice de perception de la corruption. Cet indice est construit à partir de plusieurs sondages d'opinion d'experts qui procèdent à une série d'évaluations pour plusieurs secteurs gouvernementaux, pays par pays.*

```
MortsPM          Transp          Alcool           NvDemo
Min.   : 29.00   Min.   :3.50   Min.   :0.0000   Ancienne:17
1st Qu.: 49.50   1st Qu.:4.65   1st Qu.:0.2000   Nouvelle:10
Median : 68.00   Median :6.30   Median :0.5000
Mean   : 67.67   Mean   :6.30   Mean   :0.4222
3rd Qu.: 82.50   3rd Qu.:7.90   3rd Qu.:0.5000
Max.   :112.00   Max.   :9.30   Max.   :0.9000
```

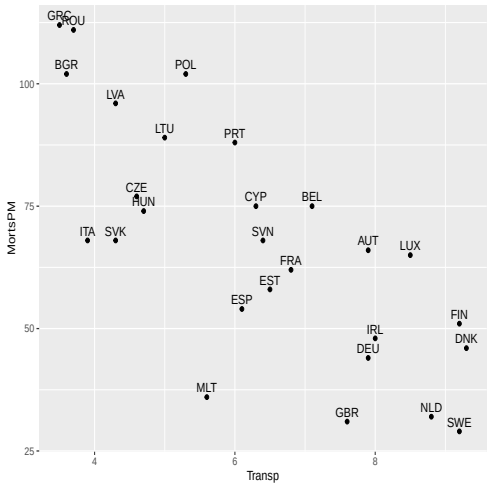
1. *Relever : la valeur de la mortalité sur les routes par million d'habitants en dessous duquel se situent 50% des pays de l'échantillon et la valeur de la mortalité sur les routes par million d'habitats au-dessus duquel se situent 25% des pays de l'échantillon.*

2. *Nous avons déterminé la matrice de corrélation.*

```
MortsPM  Transp  Alcool
MortsPM    1.000  -0.759 -0.363
Transp    -0.759   1.000  0.420
Alcool    -0.363   0.420  1.000
```

   (a) *Pourquoi n'y a-t-il pas la variable* `NuDemo` *dans le matrice de corrélation ?*

   (b) *Que peut-on dire de la corrélation linéaire entre* `MortsPM` *et* `Transp` *? On pourra également s'appuyer sur le graphique suivant qui représente le nuage de points entre la mortalité et la transparence.*



3. *On souhaite expliquer la variable* `MortsPM` *en fonction des autres variables. Nous commençons par ajuster un modèle de régression linéaire simple pour expliquer* `MortsPM` *en fonction de* `Alcool`*.*

   Modèle 1 :

```
lm(formula = MortsPM ~ Alcool, data = base)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   80.890      8.105   9.980 3.34e-10 ***
Alcool       -31.319     16.080  -1.948   0.0628 .

Residual standard error: 23 on 25 degrees of freedom
Multiple R-squared:  0.1317,Adjusted R-squared:  0.09702
F-statistic: 3.793 on 1 and 25 DF,  p-value: 0.06277
```

   (a) *Donner les estimations des coefficients de la régression. Donner l'équation de la droite ajustée.*

   (b) *Donner le coefficient de détermination et l'interpréter.*

4. *On s'intéresse désormais à l'effet de la corruption et à celui de l'ancienneté des démocraties. Pour cela on ajuste un modèle linéaire simple pour expliquer* `MortsPM` *en fonction de* `Transp` *et un autre pour expliquer* `MortsPM` *en fonction de* `NuDemo`*.*

Modèle 2 :

```
lm(formula = MortsPM ~ Transp, data = base)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  129.180     11.014   11.73 1.17e-11 ***
Transp        -9.764      1.678   -5.82 4.55e-06 ***

Residual standard error: 16.09 on 25 degrees of freedom
Multiple R-squared:  0.5753,Adjusted R-squared:  0.5584
F-statistic: 33.87 on 1 and 25 DF,  p-value: 4.549e-06
```

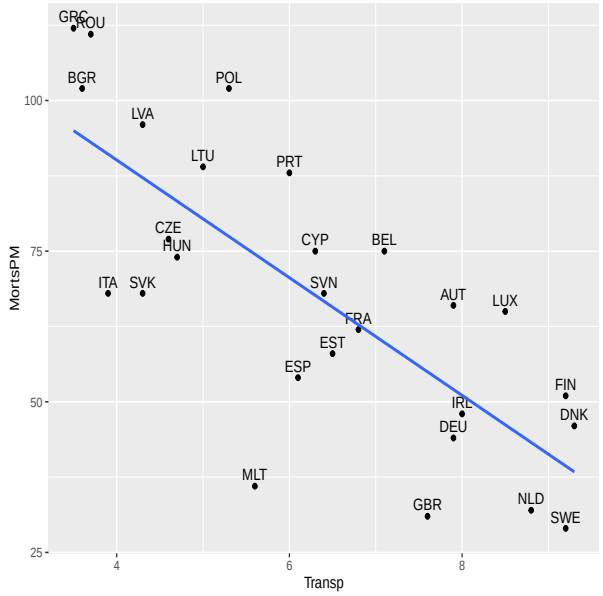
Modèle 3 :

```
lm(formula = MortsPM ~ NvDemo, data = base)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      57.765      5.026  11.492  1.8e-11 ***
NvDemoNouvelle   26.735      8.259   3.237  0.00339 **

Residual standard error: 20.72 on 25 degrees of freedom
Multiple R-squared:  0.2953,Adjusted R-squared:  0.2672
F-statistic: 10.48 on 1 and 25 DF,  p-value: 0.003393
```

(a) *Les coefficients de la régression, pour chaque modèle considéré, sont-ils significatifs au seuil 5 %? Justifier.*

(b) *Commenter* ***brièvement*** *le graphique ci-dessous.*



5. *Nous avons finalement ajusté un modèle de régression linéaire multiple expliquant* `Morts2PM` *en fonction de* `Transp2`, `Alcool` *et* `NvDemo`.

Modèle 4 :

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     118.226     19.141   6.177 2.66e-06 ***
Transp           -8.717      2.172  -4.013 0.000544 ***
Alcool            3.365     16.946   0.199 0.844329
NvDemoNouvelle    7.924     11.031   0.718 0.479775

Residual standard error: 16.54 on 23 degrees of freedom
Multiple R-squared:  0.587,Adjusted R-squared:  0.5331
F-statistic:  10.9 on 3 and 23 DF,  p-value: 0.0001191
```

(a) *Tester la significativité globale du modèle à un niveau de risque de* 5% *en n'oubliant pas de donner les hypothèses nulles et alternatives du test. Que peut-on conclure ?*

(b) *Quelles sont les variables significatives au seuil de signification de* 5% *?*

6. *Lequel des quatre modèles de régression considérés préférez vous ? Justifier. Donner l'équation du modèle ajusté que vous avez choisi et préciser l'interprétation des coefficients estimés.*