M2 SES-IES Analyse des données socio-économiques

Exercice 1

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

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

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.
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	9	10
prix observée	130.00	280.00	650.00	800.00	268.00	790.00	500.00	320.00	250.00	250.00
prix prédit	120.42	238.19	537.97	1019.75	264.96	987.64	559.38	291.72	227.49	157.89
Résidus	9.58	41.81	112.03	-219.75	3.04	-197.64	-59.38	28.28	22.51	92.11

Exercice 2

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 :

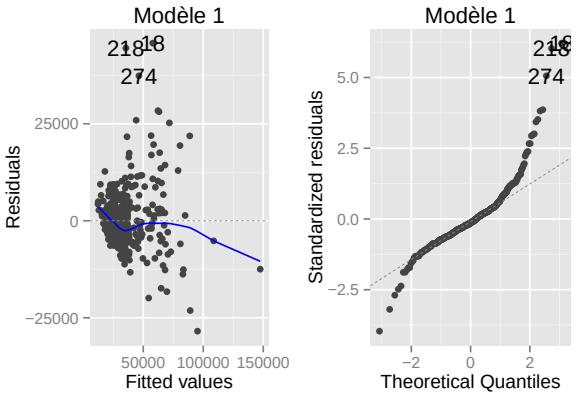
```
Modele1=lm(formula = salary ~ salbegin + jobtime + prevexp + educ + sex,
data = Salaire)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.255e+04  3.475e+03  -3.612 0.000337 ***
salbegin     1.723e+00  6.051e-02  28.472  < 2e-16 ***
jobtime      1.545e+02  3.408e+01   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) Quelles sont les variables significatives au seuil de signification de 5% ?

   (b) 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 ?

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

(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.

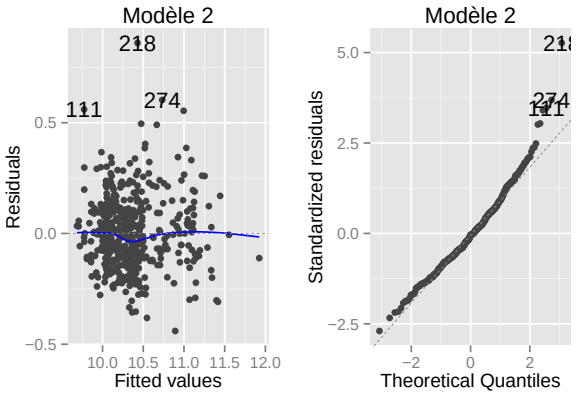
Modèle 2 :

```
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 et préciser l'interprétation des coefficients estimés.

Exercice 3

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.*

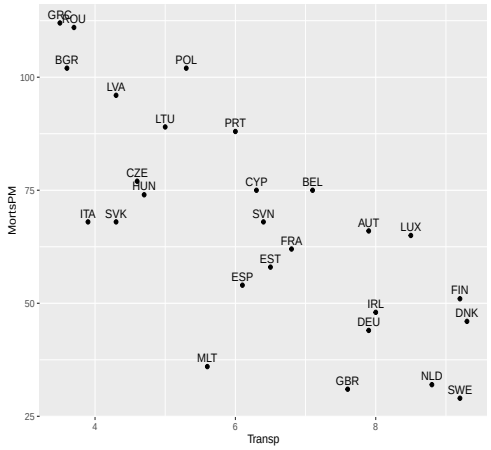
```
MortsPM          Transp         Alcool           NvDemo
Min.   : 29.00   Min.   :3.50   Min.   :0.0000   Ancienne:17
1st Qu.: 49.50   1st Qu.:4.65   1st Qu.:0.2000   Nouvelle:10
Median : 68.00   Median :6.30   Median :0.5000
Mean   : 67.67   Mean   :6.30   Mean   :0.4222
3rd Qu.: 82.50   3rd Qu.:7.90   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 `NvDemo` 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 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 et préciser leur interprétation. Donner l'équation de la droite ajustée.

(b) 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% ?

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 `NvDemo`.

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
```

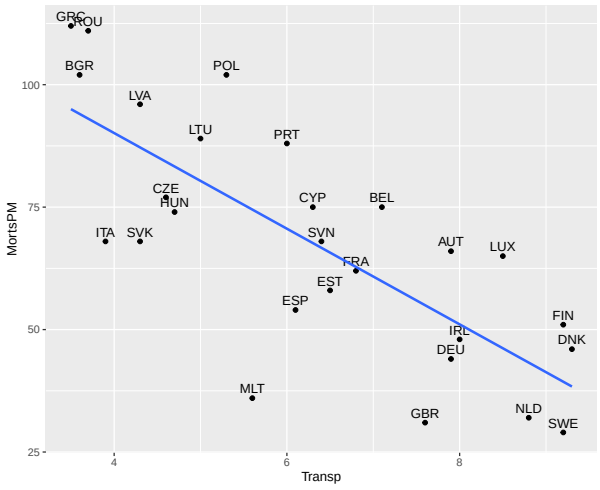
Modèle 3 :

```
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. Comment interpréter ces résultats ?

(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 :

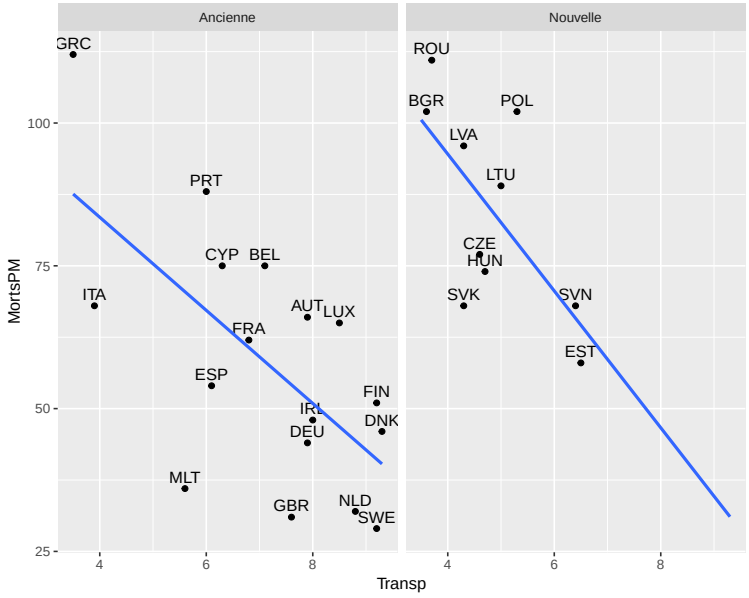
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     118.226     19.141   6.177 2.66e-06 ***
Transp           -8.717      2.172  -4.013 0.000544 ***
Alcool            3.365     16.946   0.199 0.844329
NvDemoNouvelle    7.924     11.031   0.718 0.479775

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% ?

(c) Comment expliquer que la variable `NvDemo` n'est plus utile ? On pourra s'appuyer sur le graphique suivant qui représente la mortalité en fonction de la transparence pour chacune des modalités de NvDemo.



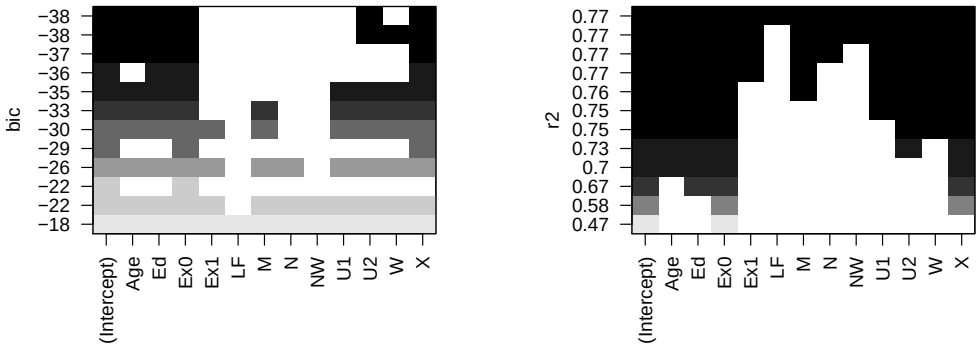
6. Lequel des quatre modèles de régression considérés préférez vous ? Justifier. Donner l'équation du modèle ajusté que vous avez choisi et préciser l'interprétation des coefficients estimés.

Exercice 4

Des criminologues américains se sont intéressés au taux de criminalité dans 47 états des Etats-Unis pendant les années 1960. Les données considérées ont été recueillies à partir d'un rapport sur la criminalité du FBI et d'autres organismes gouvernementaux afin de déterminer comment le taux de criminalité (variable à expliquer notée R) dépend d'une série de variables socio-économiques mesurées dans l'étude (taux de chômage, niveau d ?éducation, revenu médian, budget de la police,...).

- R : taux de criminalité (nombre d ?infractions signalées à la police par million d'habitants)
- Age : nombre d ?hommes de 14-24 ans pour 1000 habitants
- Ed : nombre moyen d ?années de scolarité x 10 pour les personnes de 25 ans et plus
- Ex0 : budget par habitant de la police en 1960
- Ex1 : budget par habitant de la police en 1959
- LF : taux de participation au marché du travail pour 1000 hommes âgés de 14 à 24 ans
- M : nombre d ?hommes pour 1000 femmes
- N : taille de la population de l ?État en centaines de milliers
- NW : nombre "non-whites ? ? pour 1000 habitants
- U1 : taux de chômage pour 1000 hommes âgés de 14 à 24
- U2 : taux de chômage pour 1000 hommes âgés de 35 et 39
- W : Revenu médian en dizaines
- X : nombre de familles pour 1000 gagnant moins de la moitié du revenu médian

1. On souhaite sélectionner des variables. Pour cela, on a utilisé les méthodes BIC et R^2 et obtenu les graphiques ci-dessous. Commentez et comparez ces deux graphiques. Quelles variables semblent être les plus importantes ? Quelle méthode utiliseriez-vous ? (Justifiez votre réponse).



2. On ajuste deux régression linéaires multiples

```
> rlm1 = lm(R ~ Age + Ed + Ex0 + U2 + X); summary(rlm1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -524.3743    95.1156  -5.513 2.13e-06 ***
Age            1.0198     0.3532   2.887 0.006175 **
Ed             2.0308     0.4742   4.283 0.000109 ***
Ex0            1.2331     0.1416   8.706 7.26e-11 ***
U2             0.9136     0.4341   2.105 0.041496 *
X              0.6349     0.1468   4.324 9.56e-05 ***

Residual standard error: 21.3 on 41 degrees of freedom
Multiple R-squared: 0.7296,     Adjusted R-squared: 0.6967
F-statistic: 22.13 on 5 and 41 DF,  p-value: 1.105e-10


> rlm2 = lm(R ~ Age + Ed + Ex0 + U2 + W + X); summary(rlm2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -618.5028   108.2456  -5.714 1.19e-06 ***
Age            1.1252     0.3509   3.207 0.002640 **
Ed             1.8179     0.4803   3.785 0.000505 ***
Ex0            1.0507     0.1752   5.996 4.78e-07 ***
U2             0.8282     0.4274   1.938 0.059743 .
W              0.1596     0.0939   1.699 0.097028 .
X              0.8236     0.1815   4.538 5.10e-05 ***

Residual standard error: 20.83 on 40 degrees of freedom
Multiple R-squared: 0.7478,     Adjusted R-squared:  0.71
F-statistic: 19.77 on 6 and 40 DF,  p-value: 1.441e-10
```

(a) Écrire les deux modèles théoriques considérés ici (sans oublier d'en donner les hypothèses).

(b) Dans chaque cas, tester la significativité globale du modèle à un niveau de risque $\alpha = 5\%$

(c) Interpréter le résultat du test de Fisher généraux sur R en n'oubliant pas de donner les hypothèses nulles et alternatives du test :

```
> anova(rlm1,rlm2)

Model 1: R ~ Age + Ed + Ex0 + U2 + X
Model 2: R ~ Age + Ed + Ex0 + U2 + W + X
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     41 18604
2     40 17351  1    1252.6 2.8876 0.09703 .
```

(d) Lequel des deux modèles de régression multiple considérés préférez vous ? Appuyez votre réponse à la fois sur les graphes et sur un test statistique (avec un risque $\alpha = 5\%$)

(e) Quelle(s) étude(s) supplémentaire(s) faudrait-il faire pour valider le modèle ?

Exercice 5

— Nous traitons le problème de classification de e-mails (selon une classe binaire (spam, non-spam)). On dispose de 58 variables par email.
  — Expliquer quel est le problème à traiter ?
  — Quel méthode peut-on appliquer ? Expliquer le principe de cette méthode.

— Nous avons découpé nos données en deux sous-échantillon : un échantillon d'entrainement (`Dtrain`) et un échantillon de test (`Dtest`). Nous avons ajusté un modèle logistique à l'aide du logiciel R (noté `modelFit_glm`).
  — Nous avons utilisé la fonction `predict` du logiciel R qui donne les probabilités estimées pour les deux classes pour les emails (ici le 5ème et le 7ème de `Dtest`).

```
predict(modelFit_glm, newdata=Dtest[c(5,7),],type="prob")
   nonspam      spam
   0.0006390619 0.9993609
   0.5024974029 0.4975026
```

  À quelles classes sont attribués ces emails ? Justifier.
  — Nous avons calculé la matrice de confusion sur l'échantillon de training et sur l'échantillon de test (pour un seuil s=0.5 fixé).

```
Matrice de Confusion (sur les données de Dtrain)
           Reference
Prediction nonspam spam
nonspam    1872     121
spam         80    1149

Matrice de Confusion (sur les données de Dtest)
           Reference
Prediction nonspam spam
nonspam     795   49
spam         41  494
```

  Donner le taux d'erreur sur les données de training et de test.

— Comme on a vu en cours, il vaut mieux estimer l'erreur de prédiction par validation croisée (expliquer au passage la méthode de validation croisée). *Nous avons utilisé tous nos données et nous avons estimé l'erreur de prédiction par validation croisée. L'erreur obtenue est de 0.075.*

Exercice 6

Nous traitons un problème de défaut bancaire. La variable `default` est la variable à expliquer. Nous disposons ici d'un échantillon de taille 10000 et deux variables explicatives `student` et `balance`.

— `default` : `Yes` (ou 1) si le client fait défaut sur sa dette et `No` (ou 0) sinon.
— `student` : `Yes` (ou 1) si le client est un étudiant et `No` (ou 0) sinon
— `balance` : montant moyen mensuel d'utilisation de la carte de crédit
— `income` : revenu du client.

1. On considère pour commencer un modèle de régression logistique simple où on cherche à expliquer `default` en fonction de `student`.

   (a) À l'aide du tableau de contingence ci dessous, donner les coefficients estimés du modèle logistique.

   ```
          student   No  Yes
   default
   No               6850 2817
   Yes               206  127
   ```

   (b) Nous avons utilisé le logiciel `R` pour calculer la valeur estimée du rapport de rapport de chances (odds ratio) et nous avons obtenue une valeur de 1.499. Que peut-on en conclure ?

   (c) Nous avons utilisé le logiciel R pour ajuster ce même modèle logistique simple. Donner l'équation du modèle logistique ajusté.

   Modèle 0 :

   ```
   > Modele0 = glm(formula = default~student,
                   family = binomial(link = "logit"), data = Default)
   > Modele0$coeff
   (Intercept)     balance
   -3.50413        0.40489
   ```

2. Nous avons ajusté un autre modèle logistique simple où on cherche à expliquer `default` en fonction de `balance`

   ```
   > Modele1 = glm(formula = default~student,
                 family = binomial(link = "logit"), data = Default)
   > Modele1$coeff
   (Intercept)     balance
   -10.651330614   0.005498917
   ```

   (a) Donner l'équation du modèle logistique ajusté

   (b) Nous avons relevé les valeurs estimées de la proportion de `default` pour un balance de 1000 et de 2000. À quelle classe appartient `xnew=1000` et `xnew=2000` ?

   ```
   > xnew=data.frame(balance=c(1000,2000))
   > predict.glm(Modele1,xnew,type="response")
              1           2
   0.005752145 0.585769370
   ```

3. Nous avons ajusté un modèle logistique multiple où on cherche à expliquer `default` en fonction de `student` et de `balance`.

   Modèle 2 :

```
> Modele2 = glm(formula = default ~ student + balance,
            family = binomial(link = "logit"), data = Default)
> Modele2$coeff
(Intercept)     student       balance
-10.749496    -0.714878     0.005738
```

(a) Donner l'équation du modèle logistique ajusté avec les coefficients estimés pour les `"student=Yes"` et pour les `"student=No"`.

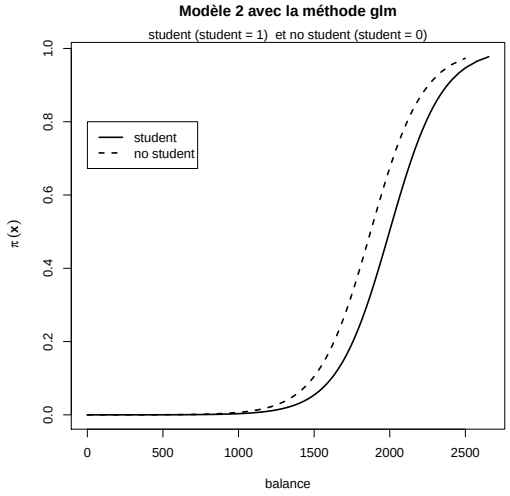
(b) Nous avons relevé les valeurs estimées de la proportion de `default` selon les caractéristiques de trois clients au hasard. Est-ce qu'on peut dire si les clients 1, 137 et 9999 feront `default` ? Justifier.

```
          1         137        9999
0.001409096 0.050602655 0.148507089
```

(c) Nous avons calculé la matrice de confusion. Donner le taux d'erreur.

```
> table(Default$default,pred.glm)
   pred.glm
       0    1
  0 9628   39
  1  228  105
```

(d) Commenter **brièvement** la figure ci-dessous.



4. Nous avons ajusté un modèle logistique complet avec toutes les variables explicatives.

```
> Modele3=glm(default~.,family = binomial(link = "logit"),  data = Default)

> Modele3$coeff
(Intercept)    student       balance       income
-1.087e+01   -6.468e-01    5.737e-03    3.033e-06
```

Donner l'équation du modèle logistique ajusté avec les coefficients estimés.

5. Nous voulons faire la sélection de modèles à l'aide du critère de AIC et de la validation croisée. Les résultats obtenus sont les suivants

— Validation croisée

```
 0.0274   0.0261  0.0269
```

— AIC

```
1600.452   1577.682   1579.545
```

Quel modèle choisir parmi les trois modèles proposés ? Justifier.