R4-L3S2 Les dépenses publiques dans l'enseignement pour chaque État

Lola LUBIN, Maëlle MARON, Ilona RATEAU et Alexis VIALATTE

2021

Table des matières

1 La base de données `education.xlsx`

1.1 Présentation

Nous allons commencer par nous poser une question, quelles sont les variables qui expliquent les dépenses publiques par personne pour l'éducation dans chaque état ?

Table 1: Résumé de la variable depenses.edu :

Moyenne	Ecart-type	Minimum	Q1	Q2	Q3	Maximum
284.600	61.340	208	234.250	269.500	316.750	546

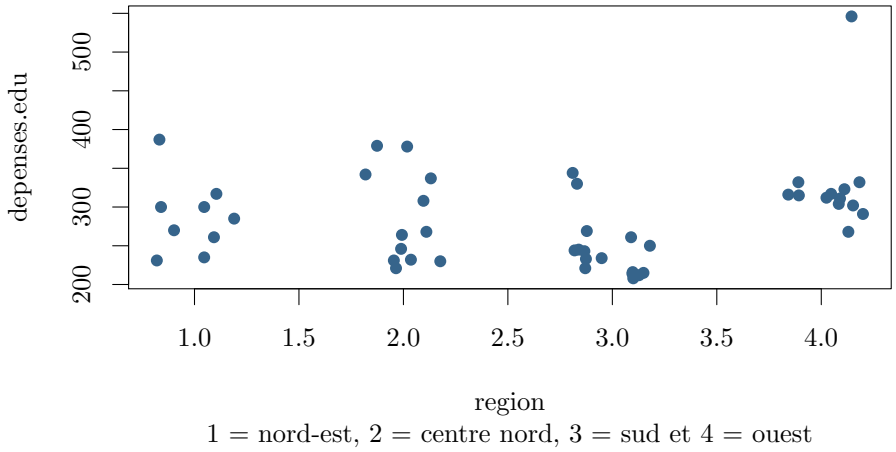
Pour répondre à cette question, nous disposons d'une base de données `education.xlsx` contenant 50 États et 6 variables. Il y a 0 donnée manquante concernant 0 État. Les variables sont :

1. etat,
2. region,
3. citadins,
4. revenu,
5. jeunes,
6. depenses.edu.

Pour chaque état, la variable `etat` correspond à son immatriculation c'est à dire AK, AL, AR, AZ, CA, CO, CT, DE, DY, FL, GA, HI, IA, ID, IL, IN, KS, LA, MA, MD, ME, MI, MN, MO, MS, MT, NB, NC, ND, NH, NJ, NM, NV, NY, OH, OK, OR, PA, RI, SC, SD, TN, TX, UT, VA, VT, WA, WI, WV, WY, la variable `region` correspond à sa « région » aux États-Unis c'est à dire 1, 2, 3, 4 où 1 = nord-est, 2 = centre nord, 3 = sud et 4 = ouest, la variable `citadins` correspond au nombre de personnes pour mille personnes vivant en ville en 1970, la variable `revenu` correspond au revenu par personne en 1973, la variable `jeunes` correspond au nombre de personnes de moins de 18 ans pour mille personnes en 1974, la variable `depenses.edu` correspond aux dépenses par personne pour l'éducation projetées en 1975.

Table 2: Aperçu de la base de données :

etat	region	citadins	revenu	jeunes	depenses.edu
LA	3	661	3825	355	244
MO	2	701	4672	309	231
TN	3	588	3946	315	212
DY	3	523	3967	325	216
RI	1	871	4780	303	300
IN	2	649	4908	329	264
GA	3	603	4243	339	250
IL	2	830	5753	320	308
SD	2	446	4296	330	230
NV	4	809	5560	330	291



Globalement, nous n'observons pas ou peu d'écart entre les régions concernant les dépenses pour l'éducation. Localement, nous observons des dépenses pour l'éducation légèrement plus élevées dans la région **ouest** que dans les autres régions et des dépenses pour l'éducation légèrement moins élevées dans la région **sud** que dans les autres régions. Les régions **nord-est** et **centre nord** sont entre les régions **sud** et **ouest**.

TABLE 3: Coefficients de corrélation linéaire de Pearson :

	citadins	revenu	jeunes	depenses.edu
citadins	1.000	0.622	-0.287	0.322
revenu	0.622	1.000	-0.297	0.608
jeunes	-0.287	-0.297	1.000	0.268
depenses.edu	0.322	0.608	0.268	1.000

1.2 La variable `citadins`

Table 4: Indépendance des variables citadins et depenses.edu :

Eff_théorique_min	p-value
0.020	0.294

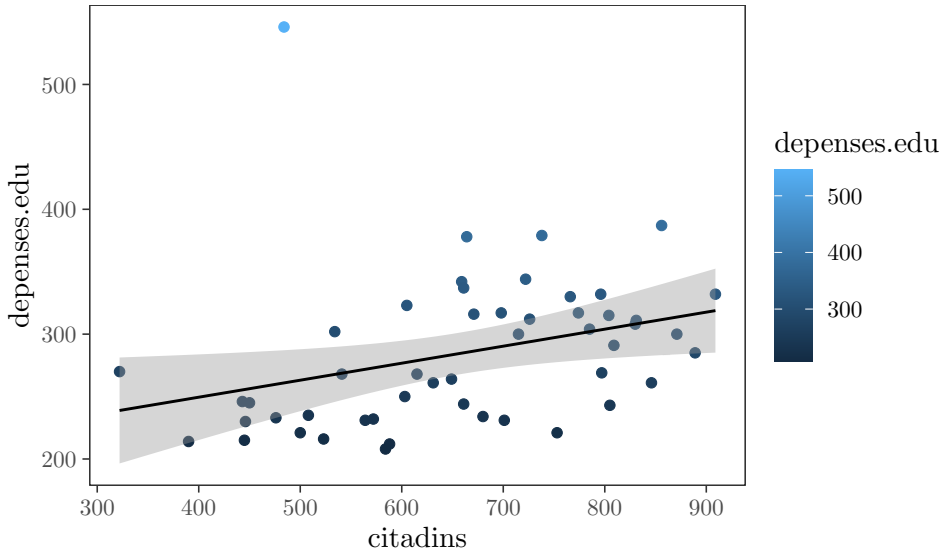
Table 5: Résumé de la variable citadins :

Moyenne	Ecart-type	Minimum	Q1	Q2	Q3	Maximum
657.800	145.016	322	546.750	662.500	782.250	909

Nous ne remplissons pas les règles nécessaires d'un test d'indépendance car le paramètre `Eff_théorique_min` est moins grand que le nombre de données de la variable `citadins`. Nous ne pouvons pas conclure que les variables `citadins` et `depenses.edu` sont indépendantes ou non. Nous supposons qu'elles sont dépendantes et que la variable `citadins` explique la variable `depenses.edu`.

Nous nous sommes demandés si plus la part de la population d'un État vivant en ville avait un impact sur les dépenses publiques par personne pour l'éducation dans ce dernier. *A priori*, nous supposons que la population en ville est – généralement – plus riche, plus éduquée, plus jeune donc elle a des enfants et elle demande à l'État de dépenser pour l'éducation de ces derniers car elle en a les moyens. Nous pensons que le coefficient de corrélation entre la variable `citadins` et la variable `depenses.edu` est strictement supérieure à 0.

Le graphique ci-dessous ne va pas à l'encontre de nos pensées et la dispersion du nuage de points a l'air corrélée de façon pertinente même si la régression linéaire simple n'est pas formelle.



1.3 La variable `revenu`

TABLE 6: Indépendance des variables revenu et depenses.edu :

Eff_théorique_min	p-value
0.020	0.254

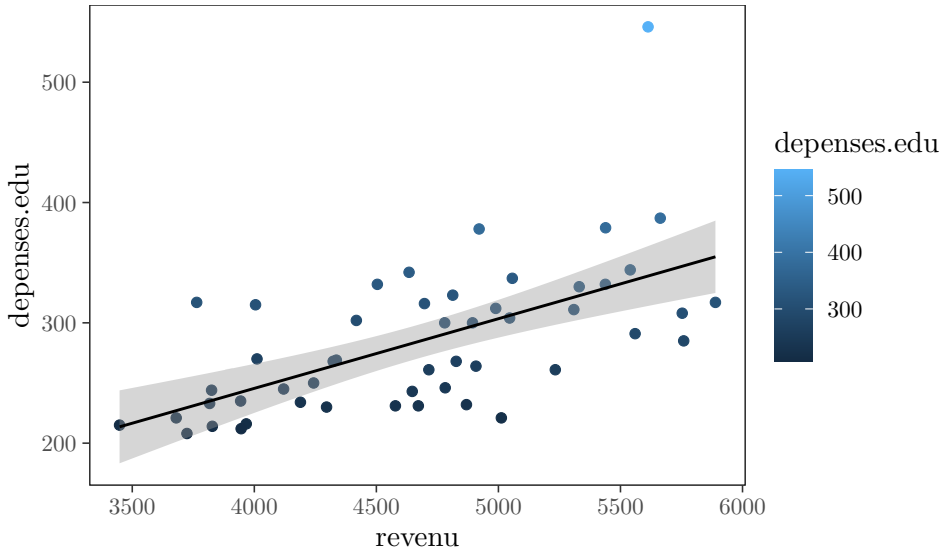
TABLE 7: Résumé de la variable revenu :

Moyenne	Ecart-type	Minimum	Q1	Q2	Q3	Maximum
4,675.120	644.506	3,448	4,137.250	4,706	5,054.250	5,889

Nous ne remplissons pas les règles nécessaires d'un test d'indépendance car le paramètre `Eff_théorique_min` est moins grand que le nombre de données de la variable `revenu`. Nous ne pouvons pas conclure que les variables `revenu` et `depenses.edu` sont indépendantes ou non. Nous supposons qu'elles sont dépendantes et que la variable `revenu` explique la variable `depenses.edu`.

Nous nous sommes demandés si plus le revenu par personne de la population d'un État avait un impact sur les dépenses publiques par personne pour l'éducation dans ce dernier. *A priori*, nous supposons que plus le revenu est élevé et plus les impôts sont élevés et donc plus l'État va dépenser pour l'éducation car il en a les moyens. Nous pensons que le coefficient de corrélation entre la variable `revenu` et la variable `depenses.edu` est strictement supérieur à 0 et plus grand que pour les deux autres variables.

Le graphique ci-dessous ne va pas à l'encontre de nos pensées et la dispersion du nuage de points a l'air corrélée de façon pertinente même si la régression linéaire simple n'est pas formelle.



1.4 La variable `jeunes`

Table 8: Indépendance des variables jeunes et depenses.edu :

Eff_théorique_min	p-value
0.020	0.369

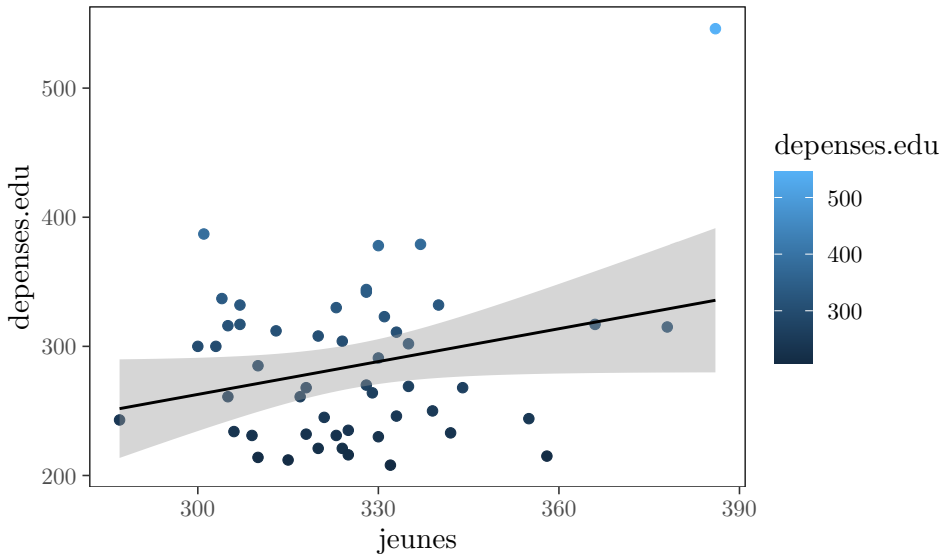
Table 9: Résumé de la variable jeunes :

Moyenne	Ecart-type	Minimum	Q1	Q2	Q3	Maximum
284.600	61.340	208	234.250	269.500	316.750	546

Nous ne remplissons pas les règles nécessaires d'un test d'indépendance car le paramètre `Eff_théorique_min` est moins grand que le nombre de données de la variable `jeunes`. Nous ne pouvons pas conclure que les variables `jeunes` et `depenses.edu` sont indépendantes ou non. Nous supposons qu'elles sont dépendantes et que la variable `jeunes` explique la variable `depenses.edu`.

Nous nous sommes demandés si plus la part de la population d'un État ayant moins de 18 ans avait un impact sur les dépenses publiques par personne pour l'éducation dans ce dernier. *A priori*, nous supposons que plus il y a de jeunes et plus l'État va dépenser pour l'éducation de ces derniers. Nous pensons que le coefficient de corrélation entre la variable `jeunes` et la variable `depenses.edu` est strictement supérieur à 0.

Le graphique ci-dessous ne va pas à l'encontre de nos pensées mais la dispersion du nuage de points n'a pas l'air corrélée de façon pertinente même si la régression linéaire simple n'est pas formelle.



2 L'analyse des régressions linéaires des dépenses pour l'éducation en fonction des citadins, du revenu et des jeunes

2.1 Présentation

D'après nos descriptions des variables explicatives ci-dessus, nous supposons $\beta_1 > 0$, $\beta_2 > 0$, $\beta_3 > 0$ et $\beta_2 > \beta_1 > \beta_3$. Nous posons :

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + z_i$$

Nous cherchons à minimiser la somme des erreurs au carré. Nous avons :

$$\text{Min}\Big(\sum_{i=1}^{50} z_i^2\Big) = \sum_{i=1}^{50}\Big[y_i - (\alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3})\Big]^2$$

2.2 Résultats

Table 10: ANOVA du modèle 1 :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
citadins	1	19129.79	19129.786	11.679	0.001
revenu	1	50042.29	50042.289	30.551	0.000
jeunes	1	39848.34	39848.343	24.328	0.000
Residuals	46	75347.58	1637.991	NA	NA

Table 11: (1/2) Régressions linéaires de l'analyse 1 :

	Dependent variable:		
	depenses.edu	log(depenses.edu)	depenses.edu
	(1)	(2)	(3)
citadins	−0.004		
	(0.051)		
revenu	0.072***		0.072***
	(0.012)		(0.009)
jeunes	1.552***		1.556***
	(0.315)		(0.308)
log(citadins)		0.056	
		(0.099)	
log(revenu)		1.052***	
		(0.171)	
log(jeunes)		1.536***	
		(0.338)	
Constant	−556.568***	−12.494***	−557.891***
	(123.195)	(2.582)	(120.864)
Observations	50	50	50
R^2	0.591	0.592	0.591
Adjusted R^2	0.565	0.565	0.574
Residual Std. Error	40.472 (df = 46)	0.130 (df = 46)	40.042 (df = 47)
F Statistic	22.186*** (df = 3; 46)	22.233*** (df = 3; 46)	33.994*** (df = 2; 47)
Note:			*p<0.1; **p<0.05; ***p<0.01

TABLE 12: (2/2) Régressions linéaires de l'analyse 1 :

	Dependent variable:		
	log(depenses.edu)	depenses.edu	log(depenses.edu)
	(1)	(2)	(3)
log(revenu)	1.107***		0.905***
	(0.140)		(0.157)
log(jeunes)	1.515***		
	(0.334)		
revenu		0.058***	
		(0.011)	
Constant	−12.472***	13.936	−2.007
	(2.563)	(51.447)	(1.325)
Observations	50	50	50
R^2	0.589	0.370	0.409
Adjusted R^2	0.572	0.357	0.397
Residual Std. Error	0.129 (df = 47)	49.191 (df = 48)	0.153 (df = 48)
F Statistic	33.677*** (df = 2; 47)	28.194*** (df = 1; 48)	33.215*** (df = 1; 48)
Note:			*p<0.1; **p<0.05; ***p<0.01

2.3 Analyse des résultats

2.3.1 Analyse des six modèles

Nous n'allons pas analyser les modèles 2, 3, 4, 5 et 6 car ils omettent des variables explicatives et ne sont pas plus pertinents que le modèle 1. Néanmoins, nous pouvons observer que la variable `revenu` est la variable explicative la plus importante de notre analyse car elle entraîne une hausse moyenne plus grande de la variable `depenses.edu` que la variable `jeunes`.

$$\mathcal{M}_1 : y_i = -556,568 - 0,004x_{i,1} + 0,072x_{i,2} + 1,552x_{i,3} + z_i$$

La constante du modèle 1 a une *p-value* du test de Student qui est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que la constante du modèle 1 est différente de 0 au seuil de 0,1%. Nous nous demandons pourquoi la constante du modèle 1 est inférieure à 0 ? Nous supposons que les préférences des États pour dépenser dans l'éducation par rapport à d'autres secteurs ne sont suffisantes tant qu'un certain palier n'est pas atteint par les variables explicatives qui sont incluses dans le modèle 1 et qui ne sont pas incluses dans le modèle 1.

Le coefficient du modèle 1 de la variable explicative `citadins` a une *p-value* du test de Student qui est égale à 0,934. Nous avons 93,4% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous conservons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative `citadins` est égal à 0. Le nombre de citadins n'a pas un impact avéré sur le modèle 1.

Le coefficient du modèle 1 de la variable explicative `revenu` a une *p-value* du test de Student qui est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative `revenu` est différent de 0 au seuil de 0,1%. Le revenu par personne a un impact avéré sur le modèle 1 c'est à dire qu'une hausse de 1$ du revenu par personne entraînera une hausse de 0,072$ des dépenses par personne pour l'éducation et inversement.

Le coefficient du modèle 1 de la variable explicative `jeunes` a une *p-value* du test de Student qui est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative

`jeunes` est différent de 0 au seuil de 0,1%. Le nombre de jeunes a un impact avéré sur le modèle 1 c'est à dire qu'une hausse du nombre de jeunes de 1 pour 1000 entraînera une hausse de 1,552$ des dépenses par personne pour l'éducation et inversement.

Le R^2 est le coefficient de détermination qui mesure la part des variables explicatives du modèle c'est à dire la précision de l'ajustement de notre modèle. Nous avons $R^2 = 0,591$ dans le modèle 1 donc 59,1% de la variation des dépenses par personne pour l'éducation est expliquée par la variation du nombre de citadins, du revenu par personne et du nombre de jeunes.

Nous avons $F = 22,186$ dans le modèle 1 et la *p-value* du test de Fisher est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse d'absence globale de significativité. Par conséquent, nous rejetons l'hypothèse d'absence globale de significativité et nous en déduisons que le modèle 1 est globalement avéré.

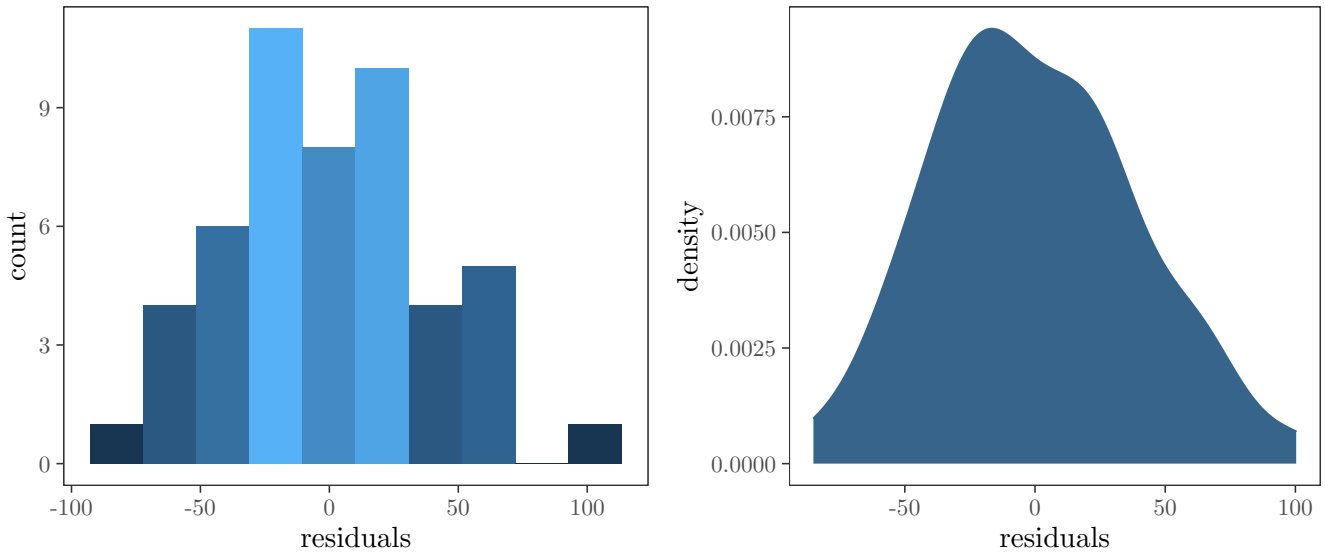
D'après nos résultats et nos analyses, nous allons conserver le modèle 3 par rapport au modèle 1 car les erreurs standards de la constante et des coefficients des variables explicatives `revenu` et `jeunes` sont moins élevées, le R^2 et le F sont globalement plus élevés et le RSE ou le $\mathcal{Z}$ est élevé. De plus, nous supprimons la variable explicative `citadins` qui s'est avérée ne pas être explicative dans notre analyse.

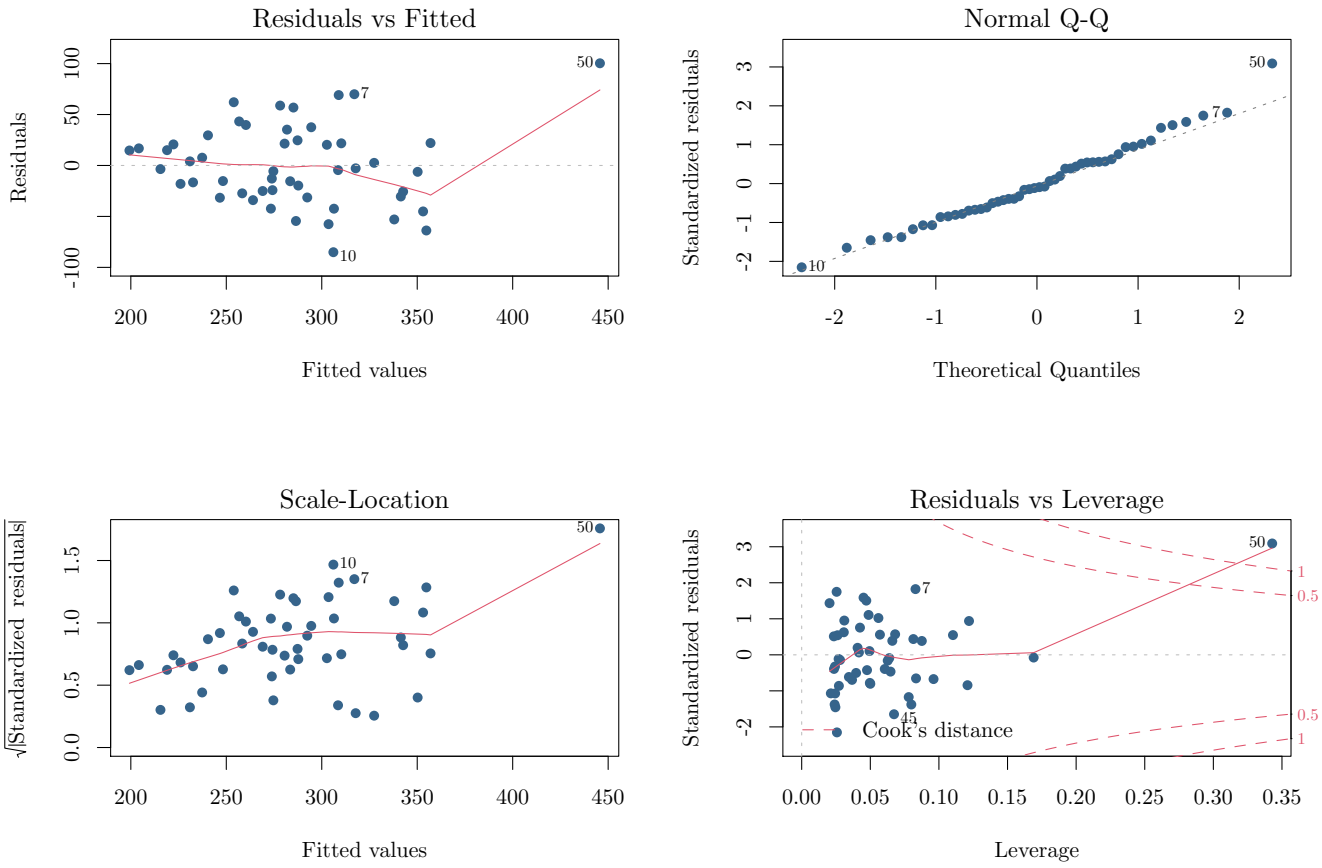
TABLE 13: ANOVA du modèle 3 :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
revenu	1	68222.10	68222.101	42.549	0
jeunes	1	40787.01	40787.015	25.438	0
Residuals	47	75358.88	1603.381	NA	NA

$$\mathcal{M}_3 : y_i = -557,891 + 0,072x_{i,2} + 1,556x_{i,3} + z_i$$

2.3.2 Analyse des résidus du modèle 3





Nous cherchons à savoir si les résidus sont distribués aléatoirement pour ne pas biaiser notre modèle. Les résidus doivent être répartis autour de la ligne horizontale représentant l'absence de résidus. Nous ne devons pas observer de tendance croissante ou décroissante. Dans notre cas, les résidus montrent une tendance globalement constante et regroupée au début mais localement croissante et dispersée à la fin donc ils ne sont pas répartis de manière globalement aléatoire. Par conséquent, grâce au test de Durbin-Watson, nous observons que les résidus ne sont pas linéaires et sont autocorrélés.

Table 14: Autocorrélation des résidus du modèle 3 :

p.value	DW
0.600	2.111

residuals

Nous cherchons à savoir si les résidus suivent une loi normale. Les résidus doivent donc être repartis le long de la droite croissante. Dans notre cas, nous observons que les résidus sont globalement répartis le long de la droite. Grâce au test de Shapiro-Wilk, nous conservons notre hypothèse car nous aurions 91,6% de chance de se tromper dans le cas contraire. Nous concluons que les résidus du modèle 3 suivent une loi normale.

Table 15: Appartenance des résidus du modèle 3 à une loi normale :

p.value	W
0.916	0.989

residuals

Nous cherchons à savoir si les résidus sont homoscédastiques. Les variances des résidus doivent donc être égales. Si les résidus sont répartis aléatoirement autour de la droite rouge et que cette dernière est horizontale alors les résidus sont répartis aléatoirement et de manière homogène autour de 0 tout au long du gradient

des valeurs estimées de la variable `depenses.edu`. Grâce au test de Harrison-McCabe, nous conservons notre hypothèse car nous aurions 90% de chance de se tromper dans le cas contraire. Nous concluons que les résidus du modèle 3 sont homoscédastiques.

TABLE 16: Homoscédasticité des résidus du modèle 3 :

p.value	HMC
0.906	0.628

residuals

Nous pouvons montrer l'effet levier des données et observer les données aberrantes grâce au quatrième graphique. Si l'effet levier est proche de 0, la suppression de la donnée n'a que peu d'incidence sur les résultats de la régression linéaire. Cependant, s'il est proche de 1 ou supérieur à 1, la suppression de la donnée a beaucup d'incidence sur les résultats de la régression linéaire. Dans notre cas, grâce à la distance de Cook, nous constatons que l'effet levier est proche de la courbe de 1 à la fin. Par conséquent, nous concluons à la présence de valeurs aberrantes.

Les trois données nous semblant aberrantes sont celles des États d'Hawaï, de l'Ohio et de New York. Nous rejetons notre hypothèse de distribution aléatoire des résidus.

2.4 Conclusion

Le modèle 3 de l'analyse 1 n'est pas bon car les hypothèses de linéarité et d'autocorrélation des résidus ne sont pas vérifiées. Nous pourrions faire mieux assez facilement en supprimant les valeurs aberrantes. Les États d'Hawaï, de l'Ohio et de New York ont des données aberrantes concernant la variable `depenses.edu` donc nous allons les supprimer de notre base de données dans l'analyse 2.

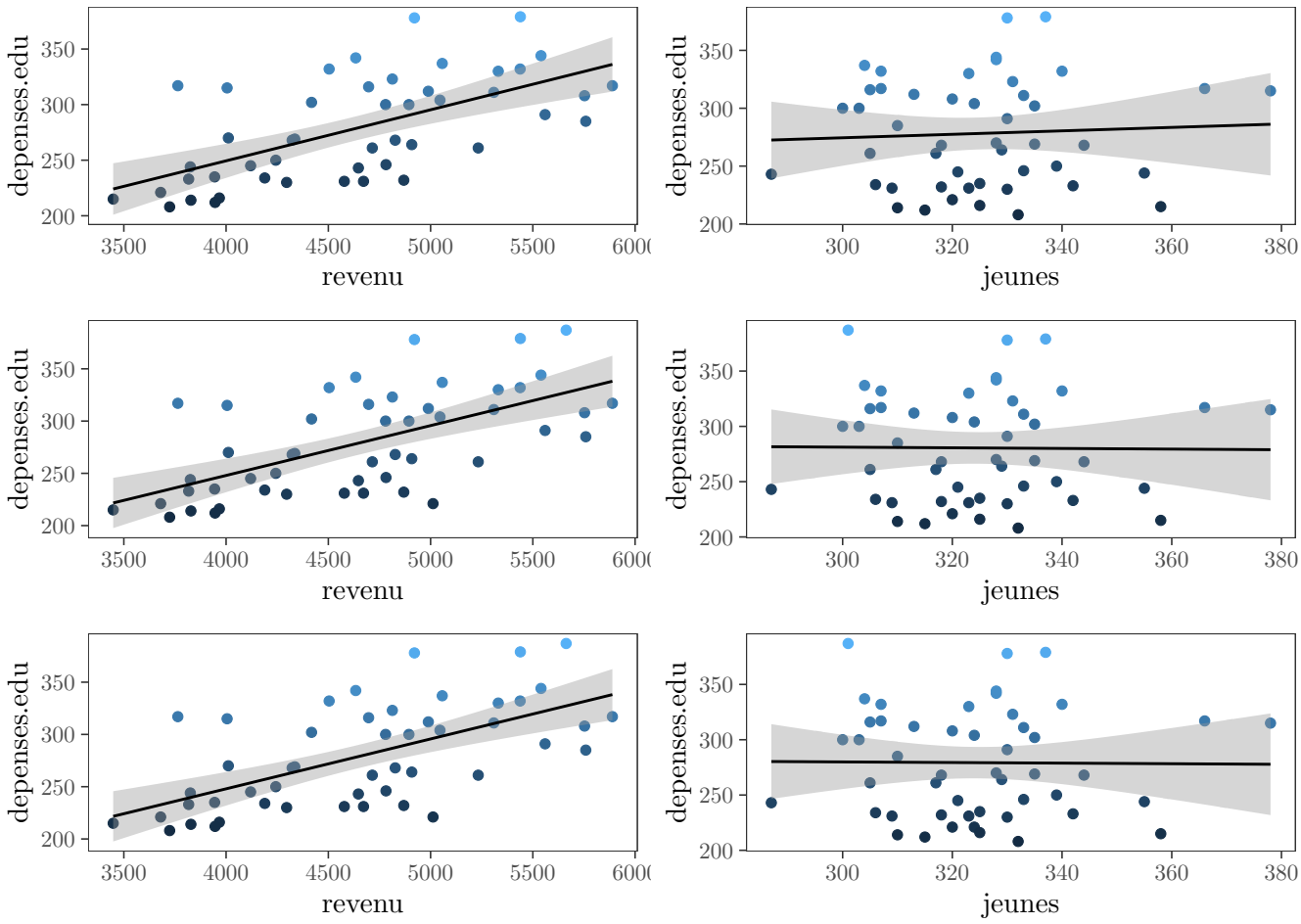
3 L'analyse de la régression linéaire des dépenses pour l'éducation en fonction du revenu et des jeunes sans les États d'Hawaï, de l'Ohio et de New York

3.1 Présentaion

Nous allons créer trois nouvelles bases de données. La première ne contiendra pas les États d'Hawaï, de l'Ohio et de New York, la deuxième ne contiendra pas les États d'Hawaï et de l'Ohio et la troisième ne contiendra pas l'État d'Hawaï. Nous allons comparer le modèle 3 de l'analyse 1 en fonction des trois nouvelles bases de données.

TABLE 17: Résumé de la variable depenses.edu

	data1	data2	data3
Moyenne	279.265	280.479	278.213
Ecart-type	48.871	48.636	46.530
Minimum	208.000	208.000	208.000
Q1	234.000	234.750	234.500
Q2	269.000	269.500	269.000
Q3	316.000	316.250	315.500
Maximum	387.000	387.000	379.000



Nos conclusions ne diffèrent pas entre ces six derniers graphiques des trois dernières bases de données et les deux premiers graphiques de la première base de données. Toutefois, nous notons cette fois ci que le coefficient de corrélation entre la variable `jeunes` et la variable `depenses.edu` est proche de 0.

3.2 Résultats

Table 18: ANOVA des modèles 1,2 et 3 :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
model1.revenu	1	44283.23	44283.228	33.923	0.000
model1.jeunes	1	10310.88	10310.882	7.899	0.007
model1.Residuals	46	60049.44	1305.423	NA	NA
model2.revenu	1	46636.92	46636.918	39.119	0.000
model2.jeunes	1	10893.01	10893.013	9.137	0.004
model2.Residuals	45	53648.05	1192.179	NA	NA
model3.revenu	1	38519.56	38519.558	34.790	0.000
model3.jeunes	1	12352.94	12352.939	11.157	0.002
model3.Residuals	44	48717.38	1107.213	NA	NA

Table 19: Intervalles de confiance au seuil de 0,05 des modèles 1, 2 et 3 :

	2.5%	97.5%
model1.Intercept	-566.448	-32.582
model1.revenu	0.041	0.078
model1.jeunes	0.265	1.603
model2.Intercept	-570.541	-59.346
model2.revenu	0.043	0.079
model2.jeunes	0.321	1.600
model3.Intercept	-571.197	-77.905
model3.revenu	0.041	0.075
model3.jeunes	0.408	1.649

Table 20: Régressions linéaires de l'analyse 2 :

	Dependent variable:		
	depenses.edu		
	(1)	(2)	(3)
revenu	0.058***	0.061***	0.059***
	(0.009)	(0.009)	(0.009)
jeunes	1.028***	0.960***	0.934***
	(0.308)	(0.318)	(0.332)
Constant	−324.551**	−314.943**	−299.515**
	(122.383)	(126.904)	(132.611)
Observations	47	48	49
R^2	0.511	0.517	0.476
Adjusted R^2	0.489	0.496	0.453
Residual Std. Error	33.275 (df = 44)	34.528 (df = 45)	36.131 (df = 46)
F Statistic	22.973*** (df = 2; 44)	24.128*** (df = 2; 45)	20.911*** (df = 2; 46)
Note:			*p<0.1; **p<0.05; ***p<0.01

3.3 Analyse des résultats

3.3.1 Analyse des trois modèles

Nous n'allons pas analyser les modèles 2 et 3 car ils ne sont pas plus pertinents que le modèle 1. Néanmoins, nous pouvons observer que la variable `revenu` est la variable explicative la plus importante de notre analyse

car car elle entraîne une hausse moyenne plus grande de la variable `depenses.edu` que la variable `jeunes`.

$$\mathcal{M}_1 : y_i = -324,551 + 0,058x_{i,2} + 1,028x_{i,3} + z_i$$

La constante du modèle 1 a une *p-value* du test de Student qui est inférieure à 0,05. Nous avons moins de 5% de chances de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que la constante du modèle 1 est différente de 0 au seuil de 5%. Nous nous demandons toujours pourquoi la constante du modèle 1 est inférieure à 0. Nous supposons que les préférences des États pour dépenser dans l'éducation par rapport à d'autres secteurs ne sont suffisantes tant qu'un certain palier n'est pas atteint par les variables explicatives qui sont incluses dans le modèle 1 et qui ne sont pas incluses dans le modèle 1.

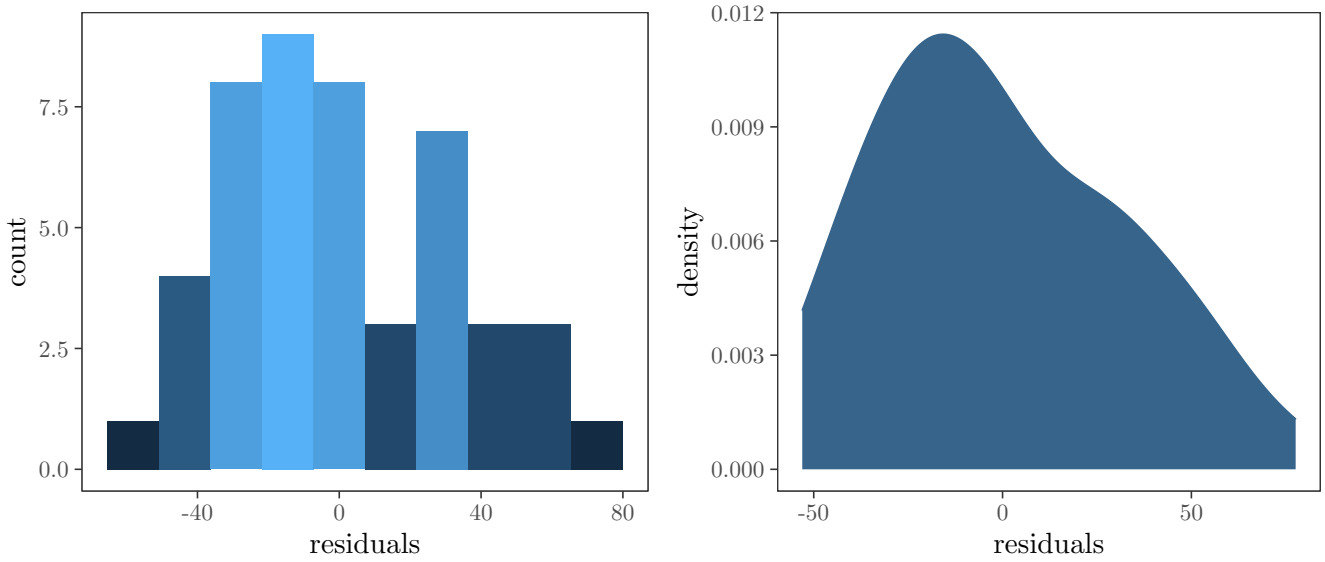
Le coefficient du modèle 1 de la variable explicative `revenu` a une *p-value* du test de Student qui est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative `revenu` est différent de 0 au seuil de 0,1%. Le revenu par personne a un impact avéré sur le modèle 1 c'est à dire qu'une hausse de 1$ du revenu par personne entraînera une hausse de 0,058$ des dépenses par personne pour l'éducation et inversement.

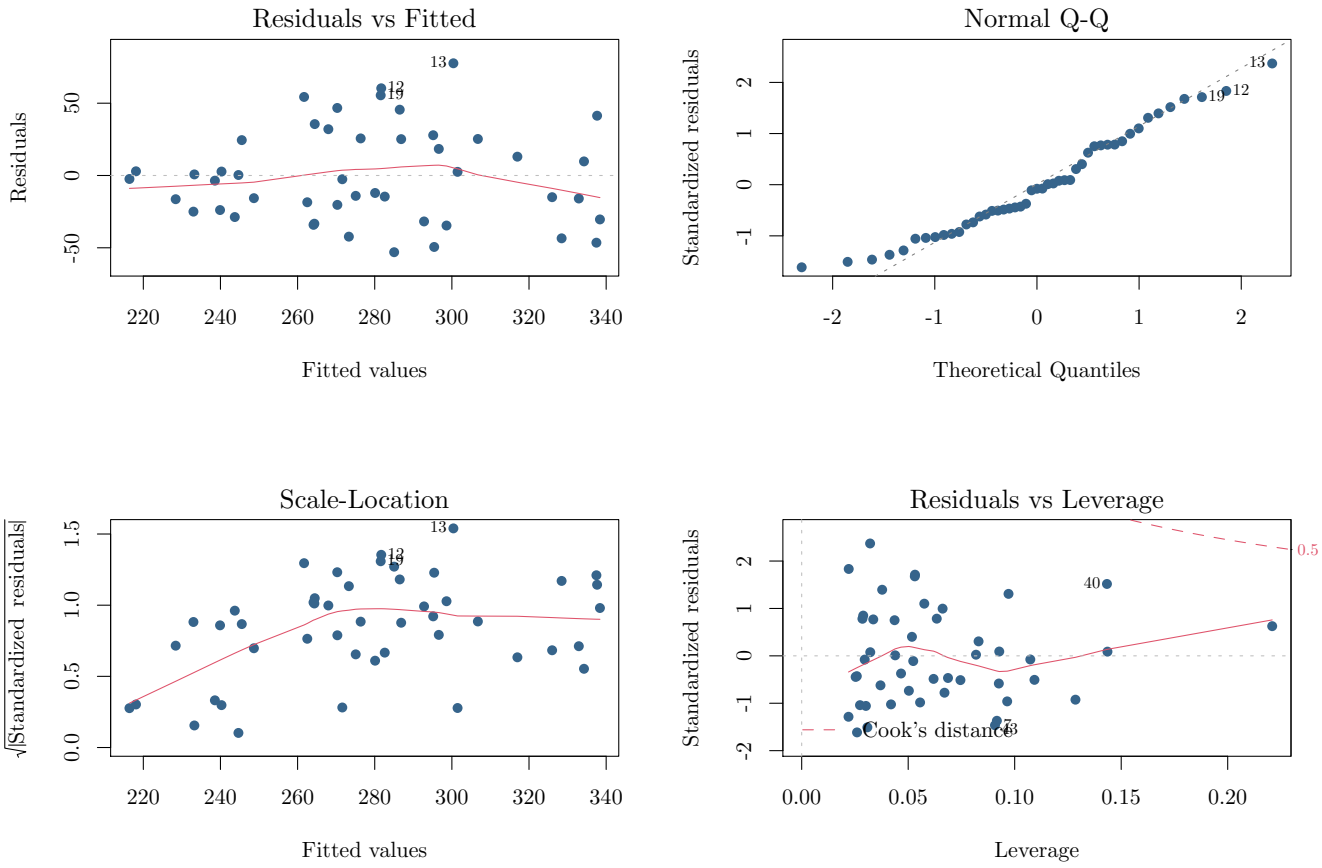
Le coefficient du modèle 1 de la variable explicative `jeunes` a une *p-value* du test de Student qui est inférieure à 0,01. Nous avons moins de 1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative `jeunes` est différent de 0 au seuil de 1%. Le nombre de jeunes a un impact avéré sur le modèle 1 c'est à dire qu'une hausse du nombre de jeunes de 1 pour 1000 entraînera une hausse de 1,028$ des dépenses par personne pour l'éducation et inversement.

Le R^2 est le coefficient de détermination qui mesure la part des variables explicatives du modèle c'est à dire la précision de l'ajustement de notre modèle. Nous avons $R^2 = 0,489$ dans le modèle 1 donc 48,9% de la variation des dépenses par personne pour l'éducation sont expliquées par la variation du revenu par personne et du nombre de jeunes.

Nous avons $F = 22,973$ dans le modèle 1 et la *p-value* du test de Fisher est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse d'absence globale de significativité. Par conséquent, nous rejetons l'hypothèse d'absence globale de significativité et nous en déduisons que le modèle 1 est globalement significatif.

3.3.2 Analyse des résidus du modèle 1





Dans notre cas, les résidus montrent une tendance globalement constante et dispersée donc ils sont répartis de manière globalement aléatoire. Par conséquent, grâce au test de Durbin-Watson, nous observons que les résidus sont linéaires mais sont autocorrélés.

Table 21: Autocorrélation des résidus du modèle 1 :

p.value	DW
0.117	1.694

residuals

Dans notre cas, nous observons que les résidus sont globalement répartis le long de la droite. Grâce au test de Shapiro-Wilk, nous conservons notre hypothèse car nous aurions 18,2% de chance de se tromper dans le cas contraire. Nous concluons que les résidus du modèle 1 suivent une loi normale.

Table 22: Appartenance des résidus du modèle 1 à une loi normale :

p.value	W
0.182	0.966

residuals

Nous cherchons à savoir si les résidus sont homoscédastiques. Grâce au test de Harrison-McCabe, nous conservons notre hypothèse car nous aurions 96,6% de chance de se tromper dans le cas contraire. Nous concluons que les résidus du modèle 1 sont homoscédastiques.

Nous pouvons montrer l'effet levier des données et observer les données aberrantes grâce au quatrième graphique. Dans notre cas, grâce à la distance de Cook, nous constatons que l'effet levier est proche de la courbe de 0,5 à la fin. Par conséquent, nous concluons à la présence de valeurs aberrantes.

Table 23: Homoscédasticité des résidus du modèle 1 :

p.value	HMC
0.958	0.670

residuals

Les trois données nous semblant aberrantes sont celles des États de l'Illinois, du Michigan et du Dakota du Sud.

3.4 Conclusion

Dans le modèle 3 de l'analyse 1 et le modèle 1 de l'analyse 2, les constantes, étant strictement inférieures à 0, nous informent que les États préfèrent dépenser ailleurs en l'abscence d'un certain niveau de la variable `revenu` et d'un certain niveau de la variable `jeunes`. La constante et les coefficient sont avérés au risque de 0,1% dans le modèle 3 de l'analyse 1 et au risque de 5% dans le modèle 1 de l'analyse 2.

La suppression des valeurs aberrantes a entraîné une baisse d'intérêt entre le modèle 1 de l'analyse 2 et le modèle 3 de l'analyse 1. De plus, nous avons perdu en précision concernant l'appartenance des résidus à une loi normale. En échange, nous avons gagné de la précision concernant l'homoscédasticité des variances ainsi que la vérification de l'hypothèse de linéarité mais pas la vérification de l'hypothèse d'autocorrélation des résidus. Étant donné que nos résultats ne divergent pas avec ou sans les valeurs aberrantes alors nos conclusions restent les mêmes.

Le modèle 1 de l'analyse 2 n'est pas bon car l'hypothèse d'autocorrélation des résidus n'est pas vérifiée. Nous pourrions faire mieux en supprimant les valeurs aberrantes. Les États de l'Illinois, du Michigan et du Dakota du Sud ont des données aberrantes concernant la variable `depenses.edu` donc nous allons les supprimer de notre base de données dans le modèle 1 de l'analyse 3.

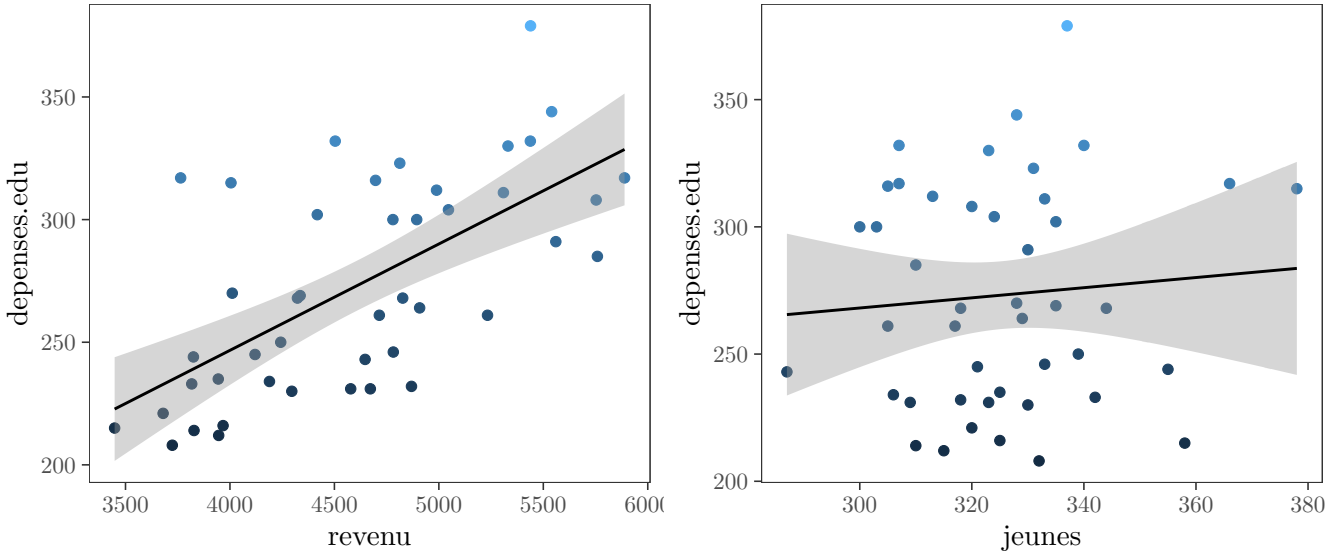
4 L'analyse de la régression linéaire des dépenses pour l'éducation en fonction du revenu et des jeunes sans les États d'Hawaï, de l'Ohio, de New York, de l'Illinois, du Michigan et du Dakota du Sud

4.1 Présentaion

Nous allons créer un nouvelle base de données qui ne contiendra pas les États d'Hawaï, de l'Ohio, de New York, de l'Illinois, du Michigan et du Dakota du Sud. Nous allons comparer le modèle 1 de l'analyse 2 en fonction de la nouvelle base de données.

Table 24: Résumé de la variable depenses.edu

Moyenne	Ecart-type	Minimum	Q1	Q2	Q3	Maximum
273.159	43.398	208	233.750	268	311.250	379



Nos conclusions ne diffèrent pas entre ces deux derniers graphiques de la dernière base de données et les six derniers graphiques des trois dernières bases de données.

4.2 Résultats

Table 25: ANOVA du modèle 1 :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
revenu	1	33838.10	33838.104	39.650	0.000
jeunes	1	12155.37	12155.374	14.243	0.001
Residuals	41	34990.41	853.425	NA	NA

Table 26: Intervalles de confiance au seuil de 0,001 du modèle 1 :

	0.05 %	99.95 %
(Intercept)	-703.746	66.501
revenu	0.028	0.082
jeunes	0.063	2.007

TABLE 27: Régression linéaire de l'analyse 3 :

	Dependent variable:
	depenses.edu
revenu	0.055^{***}
	(0.008)
jeunes	1.035^{***}
	(0.274)
Constant	-318.623^{***}
	(108.664)
Observations	44
R^2	0.568
Adjusted R^2	0.547
Residual Std. Error	29.213 (df = 41)
F Statistic	26.946^{***} (df = 2; 41)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.3 Analyse des résultats

4.3.1 Analyse du modèle

$$\mathcal{M}_1 : y_i = -318,623 + 0,055x_{i,2} + 1,035x_{i,3} + z_i$$

La constante du modèle 1 a une *p-value* du test de Student qui est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que la constante du modèle 1 est différente de 0 au seuil de 0,1%.

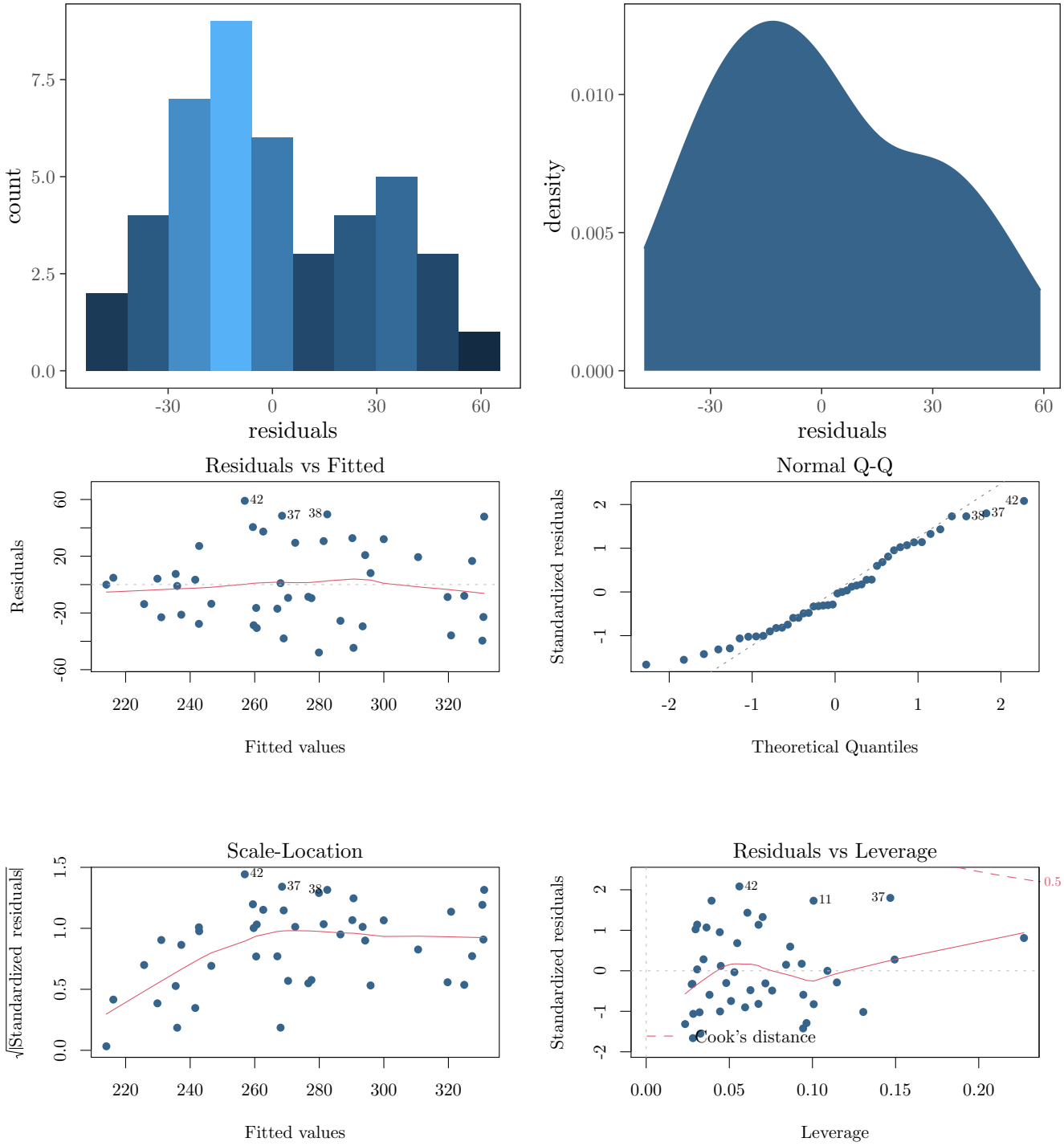
Le coefficient du modèle 1 de la variable explicative `revenu` a une *p-value* du test de Student qui est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative `revenu` est différent de 0 au seuil de 0,1%. Le revenu par personne a un impact avéré sur le modèle 1 c'est à dire qu'une hausse de 1$ du revenu par personne entraînera une hausse de 0,055$ des dépenses par personne pour l'éducation et inversement.

Le coefficient du modèle 1 de la variable explicative `jeunes` a une *p-value* du test de Student qui est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative `jeunes` est différent de 0 au seuil de 0,1%. Le nombre de jeunes a un impact avéré sur le modèle 1 c'est à dire qu'une hausse du nombre de jeunes de 1 pour 1000 entraînera une hausse de 1,035$ des dépenses par personne pour l'éducation et inversement.

Le R^2 est le coefficient de détermination qui mesure la part des variables explicatives du modèle c'est à dire la précision de l'ajustement de notre modèle. Nous avons $R^2 = 0,568$ dans le modèle 1 donc 56,8% de la variation des dépenses par personne pour l'éducation sont expliquées par la variation du revenu par personne et du nombre de jeunes.

Nous avons $F = 26,916$ dans le modèle 1 et la *p-value* du test de Fisher est inférieure à 0,001. Nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse d'absence globale de significativité. Par conséquent, nous rejetons l'hypothèse d'absence globale de significativité et nous en déduisons que le modèle 1 est globalement significatif.

4.3.2 Analyse des résidus du modèle 1



Dans notre cas, les résidus montrent une tendance globalement constante et dispersée donc ils sont répartis de manière globalement aléatoire. Par conséquent, grâce au test de Durbin-Watson, nous observons que les résidus sont linéaires mais sont autocorrélés.

Table 28: Autocorrélation des résidus du modèle 1 :

p.value	DW
0.288	1.871

residuals

Dans notre cas, nous observons que les résidus sont globalement répartis le long de la droite. Grâce au test

de Shapiro-Wilk, nous conservons notre hypothèse car nous aurions 17,5% de chance de se tromper dans le cas contraire. Nous concluons que les résidus du modèle 1 suivent une loi normale.

Table 29: Appartenance des résidus du modèle 1 à une loi normale :

p.value	W
0.175	0.963

residuals

Nous cherchons à savoir si les résidus sont homoscédastiques. Grâce au test de Harrison-McCabe, nous conservons notre hypothèse car nous aurions 56,7% de chance de se tromper dans le cas contraire. Nous concluons que les résidus du modèle 1 sont homoscédastiques.

Table 30: Homoscédasticité des résidus du modèle 1 :

p.value	HMC
0.556	0.514

residuals

Nous pouvons montrer l'effet levier des données et observer les données aberrantes grâce au quatrième graphique. Dans notre cas, grâce à la distance de Cook, nous constatons que l'effet levier est proche de la courbe de 0,5 à la fin. Par conséquent, nous concluons à la présence de valeurs aberrantes.

4.4 Conclusion

Dans le modèle 1 de l'analyse 2 et le modèle 1 de l'analyse 3, les constantes, étant strictement inférieures à 0, nous informent que les États préfèrent dépenser ailleurs en l'abscence d'un certain niveau de la variable `revenu` et d'un certain niveau de la variable `jeunes`. La constante et les coefficient sont avérés au risque de 5% dans le modèle 1 de l'analyse 2 et au risque de 0,1% dans le modèle 1 de l'analyse 3.

La suppression des valeurs aberrantes a entraîné une hausse d'intérêt entre le modèle 1 de l'analyse 3 et le modèle 1 de l'analyse 2. Néanmoins, nous avons perdu en précision concernant l'appartenance des résidus à une loi normale et l'homoscédasticité des variances. De plus, nous n'avons pas gagné la vérification de l'hypothèse d'autocorrélation des résidus.

5 Synthèse

Nous avons remarqué que, pour chaque État, le revenu par personne est ce qui influence le plus les dépenses publiques par personne pour l'éducation. Nous pensons que dans ce cas, les États en ont les moyens. Nous avons aussi remarqué que le nombre de jeunes de moins de 18 ans pour mille personnes influence aussi les dépenses publiques par personne pour l'éducation car dans ce cas, les États en ont le besoin.

Nous ne pensons pas pouvoir vérifier l'hypothèse d'autocorrélation des résidus car nous risquerions de ne plus vérifier l'hypothèse d'appartenance des résidus à une loi normale...

6 tools

Nous nous sommes servis de ces tools :

```
tool.table <- function( b, c, d ){
  a <- b %>%
    kable( caption = c,
           digits = 3,
           booktabs = T ) %>%
    kable_styling( full_width = F,
                   position = "center",
                   latex_options = c( "striped",
                                      "condensed",
                                      "hold_position",
                                      d ) )
  return( a )
}# pour les tableaux,
# a = ( b = data, c = caption, d = "scale_down" ),

tool.summary <- function( b ){
  a <- c( mean( b,
                na.rm = T ),
          sd( b,
              na.rm = T ),
          quantile( b,
                    na.rm = T ) )
  a <- round( a,
              3 )
  names( a ) <- c( "Moyenne",
                   "Ecart-type",
                   "Minimum",
                   "Q1",
                   "Q2",
                   "Q3",
                   "Maximum" )
  return( a )
}# pour les résumés,
# a = ( b = data ),

tool.durbin <- function( b ){
  test1 <- dwtest( b )
  a <- c( test1$p.value,
          test1$statistic )
  a <- round( a,
              3 )
  names( a ) <- c( "p.value",
                   "DW" )
  return ( a )
}# pour les tests de Durbin-Watson,
# a = ( b = data ),

tool.shapiro <- function( b ){
  test1 <- shapiro.test( b )
  a <- c( test1$p.value,
```

```
          test1$statistic )
  a <- round( a,
              3 )
  names( a ) <- c( "p.value",
                   "W" )
  return( a )
}# pour les tests de Shapiro-Wilk,
# a = ( b = data ),

tool.harrison <- function( b ){
  test1 <- hmctest( b )
  a <- c( test1$p.value,
          test1$statistic )
  a <- round( a,
              3 )
  names( a ) <- c( "p.value",
                   "HMC" )
  return ( a )
}# pour les tests de Harrison-McCabe,
# a = ( b = data ),

tool.student <- function( b, c ){
 test1 <- var.test( b ~ c )
 test2 <- t.test( b ~ c,
                  equal = test1$p.value > 0.05 )
 a <- c( table( c[ !is.na( b ) ] ),
         ifelse( test2$p.value > 0.05,
                 "Oui",
                 "Non" ),
         round( c( test2$estimate,
                   test2$p.value ),
                3 ) )
 names( a ) <- c( names( table( c[ !is.na(b) ] ) ),
                  "var.equal",
                  names( test2$estimate ),
                  "p.value" )
 return( a )
}# pour les tests de Student de comparaison des moyennes et des variances,
# a = ( b = data1, c = data2 ),

tool.chi2 <- function( b, c ){
  test1 <- chisq.test( b,
                       c )
  a <- c( min( test1$expected ),
          test1$p.value )
  a <- round( a,
              3 )
  names( a ) <- c( "Eff_théorique_min",
                   "p-value" )
  return( a )
}# pour les tests du Chi 2 d'indépendance des variances,
# a = ( b = data1, c = data2 ),
```

7 packages

Nous nous sommes servis de ces packages :

```
library( readxl )# pour la base de données,
library( dplyr )# pour la syntaxe,
library( tidyverse )
library( forcats )# pour les vecteurs,
library( lmtest )# pour les tests,

library( plm )# pour les régressions,
library( car )# Pour les régressions,
library( carData )# Pour les régressions,
library( stargazer )# pour les régressions,
library( lmtest )# pour les régressions,
library( statsr )# pour les régressions,
# library( summarytools )# pour les régressions,

library( printr )# pour les tableaux,
library( knitr )# pour les tableaux,
library( kableExtra )# pour les tableaux,
library( modelsummary )# pour les tableaux,
library( gtsummary )# pour les tableaux,

# library( ggplot1 )# pour les graphiques,
library( ggplot2 )# pour les graphiques,
library( ggcorrplot )# pour les graphiques,
library( ggfortify )# pour les graphiques,
library( ggpubr )# pour les graphiques,
library( ggrepel )# pour les graphiques,
library( ggridges )# pour les graphiques,
library( ggsci )# pour les graphiques,
library( ggsignif )# pour les graphiques,
```