2024-10-20

Devoir Individuel

Analyse et Inférence Statistique (MATH 60619)

Antoine Laverdière-Allaire (11306200)
HEC MONTRÉAL

Question 1 : Présentez des statistiques descriptives appropriées pour toutes les variables disponibles.

Nous disposons de cinq variables pour cette étude : revenge_t1, revenge_t5, age, vc (vindictive complaining) et wom (negative word-of-mouth).

```
PROC MEANS data=infe.Devoirl n mean std min max median Q1 Q3;
  var revenge_tl revenge_t5 age vc wom;
RUN;
```

À l'aide de ce code et de la procédure Means, on est en mesure d'obtenir des statistiques descriptives appropriées pour ces variables comme la taille de l'échantillon (N), la moyenne, la médiane, l'écart-type, les valeurs minimales et maximales, ainsi que les quartiles.

Voici les résultats :

Le Système SAS

La procédure MEANS

Variable	N	Moyenne	Ec-type	Minimum	Maximum	Médiane	Quartile inférieur	Quartile supérieur
revenge_t1	80	4.5600000	1.3241286	1.2000000	7.0000000	4.6000000	3.6000000	5.6000000
revenge_t5	80	2.2825000	1.1136410	1.0000000	5.6000000	2.1000000	1.3000000	3.1000000
age	80	42.0750000	7.4863166	22.0000000	61.0000000	43.0000000	38.0000000	47.0000000
vc	80	3.2781250	1.7668689	1.0000000	7.0000000	2.8750000	1.7500000	4.6250000
wom	80	4.0083333	1.4082088	1.0000000	7.0000000	3.8333333	3.0000000	5.0000000

1. Revenge_t1 (Désir de vengeance à la première vague) :

Le score moyen de vengeance lors de la première vague est de 4,56 avec un écart-type de 1,32, indiquant une dispersion modérée autour de la moyenne. La médiane de 4,60 est très proche de la moyenne et suggère une distribution symétrique. La majorité des scores se situent entre 3,60 (Q1) et 5,60 (Q3), indiquant une répartition centrale des scores avec des valeurs autour de la moyenne.

2. Revenge_t5 (Désir de vengeance à la cinquième vague) :

À la cinquième vague, le score moyen de vengeance a considérablement diminué à 2,28, suggérant une baisse du désir de vengeance. L'écart-type est plus faible à 1,11, montrant une dispersion légèrement plus faible qu'au temps 1 et signifiant que les réponses des participants sont plus homogènes. La médiane de 2,1 est légèrement inférieure à la moyenne, indiquant une légère asymétrie vers la droite. La moitié des scores sont compris entre 1,30 (Q1) et 3,10 (Q3)

3. Âge des participants :

L'âge des participants varie entre 22 et 61 ans, avec une moyenne de 42,08 ans. L'écart-type est de 7,49 ans. La médiane est de 43 ans et très proche de la moyenne, indiquant une distribution assez symétrique avec une majorité des participants âgés entre 38 (Q1) et 47 (Q3) ans.

4. Vindictive Complaining (vc) :

Le comportement de plainte vindicative a une moyenne de 3,28 avec un écart-type de 1,77, démontrant une grande variabilité de la distribution. La médiane de 2,88 est inférieure à la moyenne, suggérant encore une asymétrie vers la droite. La moitié des scores sont compris entre 1,75 (Q1) et 4,625 (Q3).

5. Negative Word-of-Mouth (wom):

La moyenne de wom est de 4,01 avec un écart-type de 1,41, montrant une dispersion modérée. La médiane de 3,83 est légèrement inférieure à la moyenne et indique une asymétrie modérée vers la droite. La moitié des participants ont des scores compris entre 3,00 (Q1) et 5,00 (Q3).

Conclusion :

Les statistiques descriptives montrent que le désir de vengeance moyen des participants s'estompe au fil du temps (de la première à la cinquième vague). L'âge des participants est relativement homogène, tandis que les comportements de plainte vindicative et de diffusion de bouche-à-oreille négatif varient davantage. La médiane et les quartiles ajoutent une perspective utile, notamment pour évaluer la symétrie des distributions et identifier les différences de comportements.

Question 2 : Est-ce qu'il y a une différence entre le score de vengeance moyen des hommes et des femmes à la première vague?

Pour répondre à cette question, on utilise la procédure TTEST dans SAS pour comparer les moyennes entre les deux groupes à la première vague : le score de vengeance des hommes et le score de vengeance des femmes. On cherche donc à tester l'hypothèse nulle suivante : H0 : μ1 = μ2

où μ1 = score de vengeance moyen pour les hommes et μ2 = score de vengeance moyen pour les femmes

```
PROC TTEST data=infe.Devoir1;
   class sexe;
   var revenge_t1;
RUN;
```

Voici les résultats :

La procédure TTEST

Variable : revenge_t1

sexe	Méthode	N	Moyenne	Ec-type	Err. type	Minimum	Maximum
0		42	4.2619	1.3888	0.2143	1.2000	7.0000
1		38	4.8895	1.1807	0.1915	2.4000	7.0000
Diff (1-2)	Pooled		-0.6276	1.2942	0.2898		
Diff (1-2)	Satterthwaite		-0.6276		0.2874		

sexe	Méthode	Moyenne	IC à95% - Moyenne		Ec-type	Ec.-type de l'IC à95%	
0		4.2619	3.8291	4.6947	1.3888	1.1427	1.7709
1		4.8895	4.5014	5.2775	1.1807	0.9625	1.5275
Diff (1-2)	Pooled	-0.6276	-1.2044	-0.0507	1.2942	1.1191	1.5348
Diff (1-2)	Satterthwaite	-0.6276	-1.1998	-0.0553			

Méthode	Variances	DDL	Valeur du test t	Pr > |t|
Pooled	Egal	78	-2.17	0.0334
Satterthwaite	Non égal	77.714	-2.18	0.0320

Egalité des variances

Méthode	DDL num.	DDL den.	Valeur F	Pr > F
Folded F	41	37	1.38	0.3194

Les résultats du test de différence entre les scores moyens de vengeance à la première vague (Revenge_t1) pour les hommes (sexe = 0) et les femmes (sexe = 1) sont significativement différents. En effet, la moyenne des hommes est de 4,2619, alors que la moyenne des femmes se situe à 4,8895. Cela implique une différence des moyennes de -0,6276.

Avant de continuer avec l'interprétation des résultats, il faut effectuer le test de l'égalité des variances pour savoir quelle méthode choisir.

Egalité des variances				
Méthode	DDL num.	DDL den.	Valeur F	Pr > F
Folded F	41	37	1.38	0.3194

Dans le tableau, on aperçoit une p-value de 0,3194. Cette valeur est supérieure à 0,05, ce qui signifie que nous ne rejetons pas l'hypothèse nulle d'égalité des variances. En d'autres mots, les variances entre les groupes d'homme et de femmes ne sont pas statistiquement différentes à un seuil de 5%. Les variances sont donc considérées comme étant égales et on utilise la version Pooled.

Méthode	Variances	DDL	Valeur du test t	Pr > |t|
Pooled	Egal	78	-2.17	0.0334
Satterthwaite	Non égal	77.714	-2.18	0.0320

Ensuite, lorsqu'on regarde la p-value associée au test-t de la méthode Pooled (0,0334), on s'aperçoit qu'elle est inférieure à 0,05. Cela signifie qu'on va rejeter l'hypothèse nulle initiale au seuil de 5%. On peut donc conclure qu'il existe une différence statistiquement significative entre les hommes et les femmes en ce qui concerne leur score de vengeance à la première vague.

sexe	Méthode	Moyenne	IC à95% - Moyenne		Ec-type	Ec.-type de l'IC à95%	
0		4.2619	3.8291	4.6947	1.3888	1.1427	1.7709
1		4.8895	4.5014	5.2775	1.1807	0.9625	1.5275
Diff (1-2)	Pooled	-0.6276	-1.2044	-0.0507	1.2942	1.1191	1.5348
Diff (1-2)	Satterthwaite	-0.6276	-1.1998	-0.0553			

Les intervalles de confiance à 95% pour la différence des moyennes sont de -1,2044 à -0,0507. Cela signifie qu'avec un niveau de confiance de 95%, la véritable différence entre les moyennes des scores de vengeance des hommes et des femmes se situe probablement entre ces valeurs. Comme l'intervalle ne contient pas zéro, cela confirme la significativité statistique de la différence.

L'analyse montre donc que les femmes ont un score de vengeance significativement plus élevé en moyenne que les hommes lors de la première vague (4,8895 contre 4,2619). La p-value de la méthode Pooled du test de Welch et l'intervalle de confiance à 95% confirment tous deux qu'il existe une différence statistiquement significative entre les deux groupes.

Question 3 : Est-ce que le désir de vengeance a tendance à s'estomper dans le temps?

Pour répondre à cette question, on utilise la procédure TTEST dans SAS afin de comparer la différence entre les scores de vengeance à la première vague (revenge_t1) et à la cinquième vague (revenge_t5).

On cherche donc à tester l'hypothèse nulle suivante : $H0 : \mu1 = \mu2$

où $\mu1$ = score de vengeance moyen mesuré à la première vague (revenge_t1) et
$\mu2$ = score de vengeance moyen mesuré à la cinquième vague (revenge_t5)

```
17  /* Question 3 */
18
19  PROC TTEST data=infe.Devoir1;
20    paired revenge_t1*revenge_t5;
21  RUN;
22
```

Le Système SAS

La procédure TTEST

Différence : revenge_t1 - revenge_t5

N	Moyenne	Ec-type	Err. type	Minimum	Maximum
80	2.2775	0.7507	0.0839	0.2000	4.0000

Moyenne	IC à95% - Moyenne		Ec-type	Ec.-type de l'IC à95%	
2.2775	2.1104	2.4446	0.7507	0.6497	0.8892

DDL	Valeur du test t	Pr > |t|
79	27.14	< .0001

Tout d'abord, on aperçoit que la différence moyenne entre revenge_t1 et revenge_t5 est de 2,2775. Cela signifie qu'en moyenne, le désir de vengeance des participants diminue de 2,2775 points entre la première et la cinquième vague, indiquant une diminution du désir de vengeance avec le temps.

L'intervalle de confiance à 95% pour la moyenne de la différence est [2,1104 ; 2,4446]. Cela signifie qu'avec un niveau de confiance de 95%, la différence entre le désir de vengeance mesuré à la première vague et celui mesuré à la cinquième vague se situe probablement entre 2,1104 et 2,4446 points sur l'échelle de 1 à 7. Cet intervalle ne contient pas zéro suggérant qu'il est extrêmement improbable (à 95 %) que la différence observée soit due au hasard. Cela confirme aussi que le désir de vengeance a significativement diminué entre les deux vagues, car la différence observée est positive ce qui montre une réduction du score de vengeance.

Le test t donne une valeur de 27,14 avec une p-value < 0.0001. Il s'agit donc d'une différence importante comparé à ce qu'on s'attendrait si l'hypothèse nulle (H0) était vraie. Cela signifie aussi que la probabilité que la différence entre revenge_t1 et revenge_t5 soit due au hasard est extrêmement faible (moins de 0,01%). Puisque la p-value est très petite (< 0.05), on rejette l'hypothèse nulle selon laquelle il n'y aurait pas de différence entre les deux vagues. En d'autres mots, la diminution du désir de vengeance est statistiquement significative.

Pour conclure, les résultats montrent clairement que le désir de vengeance s'estompe effectivement dans le temps et que les participants sont moins enclins à vouloir se venger à la cinquième vague.

Question 4 : Quelle est la variable la plus corrélée à la variable vengeance au temps 1? Interprétez le résultat

Pour répondre à cette question, on utilise la procédure CORR dans SAS afin de comparer les coefficients de corrélation de Pearson entre la variable revenge_t1 (score de vengeance au temps 1) et les trois autres variables à analyser : age, vc (vindictive complaining) et wom (negative word-of-mouth). On n'inclut pas la variable sexe puisqu'il s'agit d'une variable catégorielle et non adaptée pour la corrélation de Pearson.

Voici le code SAS utilisé afin d'obtenir les résultats :

```
25 PROC CORR data=infe.Devoir1;
26    var revenge_t1 age vc wom;
27 RUN;
```

Voici les résultats obtenus dans SAS :

4 Variables : revenge_t1 age vc wom

Statistiques simples

Variable	N	Moyenne	Ec-type	Somme	Minimum	Maximum
revenge_t1	80	4.56000	1.32413	364.80000	1.20000	7.00000
age	80	42.07500	7.48632	3366	22.00000	61.00000
vc	80	3.27813	1.76687	262.25000	1.00000	7.00000
wom	80	4.00833	1.40821	320.66667	1.00000	7.00000

Coefficients de corrélation de Pearson, N = 80
Proba > |r| sous H0: Rho=0

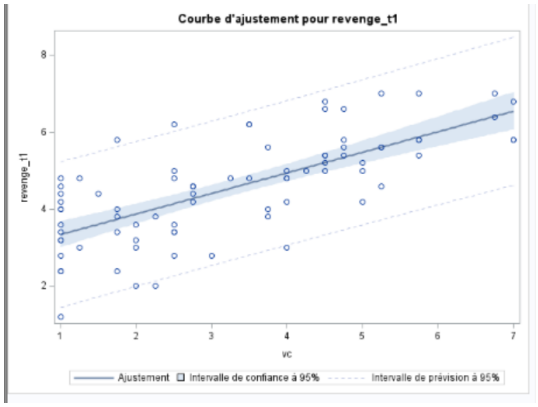
	revenge_t1	age	vc	wom
revenge_t1	1.00000	0.46793 <.0001	0.71251 <.0001	0.28937 0.0092
age	0.46793 <.0001	1.00000	0.38478 0.0004	-0.17817 0.1138
vc	0.71251 <.0001	0.38478 0.0004	1.00000	-0.26422 0.0179
wom	0.28937 0.0092	-0.17817 0.1138	-0.26422 0.0179	1.00000

Comme on peut le constater, la variable la plus fortement corrélée avec le désir de vengeance au temps 1 (revenge_t1) est vc (vindictive complaining) avec un coefficient de corrélation de 0,71251. Puisque la corrélation entre revenge_t1 et vc est proche de 1, les points de la distribution ont plus tendance à être alignés autour d'une droite de pente positive. Par conséquent, plus les participants ont un comportement de plainte vindicative (vc) qui est élevé, plus leur score de vengeance au temps 1 (revenge_t1) aura tendance a être élevé aussi.

Il existe donc une forte corrélation positive entre les deux variables et la variable vc est celle qui est la plus étroitement lié à la variable revenge_t1. Avec une p-value inférieure à 0.0001, on remarque aussi que cette corrélation est statistiquement significative.

Afin de s'assurer que le coefficient de corrélation est bien interprété, on va examiner le diagramme de dispersion des deux variables à l'aide du code SAS ci-dessous.

```
ods rtf;
PROC GLM data=infe.Devoir1;
   model revenge_t1 = vc;
RUN;
```



On s'aperçoit en analysant le diagramme de dispersion que la valeur du coefficient de corrélation linéaire de Pearson interprète bien le lien fort entre les deux variables et la distribution des valeurs qui se situe autour de la droite de pente positive.

Question 5 : Est-ce que le sexe a un effet significatif sur le désir de vengeance (au temps 1) une fois que l'on tient compte de l'âge?

Pour répondre à cette question, on utilise la procédure REG dans SAS pour effectuer une régression linéaire multiple afin d'évaluer l'impact simultané des deux variables explicatives (sexe et âge) sur la variable dépendante (revenge_t1). Nous voulons tester si le sexe a un effet significatif sur revenge (t_1) tout en contrôlant l'âge. Tenir compte de l'âge signifie d'examiner l'effet du sexe sur revenge_t1 indépendamment de l'âge et ainsi d'interpréter le coefficient β de la même manière que la régression simple.

L'hypothèse nulle est donc : H0 : β sexe = 0 (où le sexe n'a pas d'effet sur revenge_t1)

Voici le code SAS utilisé afin d'obtenir les résultats obtenus ci-dessous :

```
PROC REG data=infe.Devoir1;
   model revenge_t1 = sexe age;
RUN;
```

Le Système SAS

La procédure REG
Modèle : MODEL1
Variable dépendante : revenge_t1

Nb d'observations lues	80
Nb d'obs. utilisées	80

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	2	39.59393	19.79696	15.41	<.0001
Erreur	77	98.91807	1.28465		
Total sommes corrigées	79	138.51200			

Root MSE	1.13342	R carré	0.2859
Moyenne dépendante	4.56000	R car. ajust.	0.2673
Coeff Var	24.85579		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr> |t|
Intercept	1	0.67040	0.74345	0.90	0.3700
sexe	1	0.68215	0.25400	2.69	0.0089
age	1	0.08474	0.01705	4.97	<.0001

La valeur du coefficient de régression pour la variable sexe est de 0,68215. Cela signifie que, toutes choses étant égales par ailleurs, revenge_t1 est en moyenne 0,68215 points plus élevés pour les femmes (sexe = 1) que pour les hommes (sexe = 0). En effet, toutes choses étant égales par ailleurs et une fois l'âge contrôlé, lorsque le sexe passe de 0 à 1, revenge_t1 augmente de 0,68215 points en moyenne. La p-value associée à la variable sexe est de 0,0089 et inférieure au seuil de significativité de 0,05. On va donc rejeter l'hypothèse nulle (H0).

On conclut donc, après avoir effectué la régression linéaire multiple, que le sexe a un effet significatif sur le score de vengeance au temps 1, même après avoir pris en compte l'effet de l'âge.

<u>Question 6 : Est-ce que le sexe a un effet significatif sur le désir de vengeance (au temps 1) une fois que l'on tient compte de toutes les variables explicatives mesurées au temps 1?</u>

a) Présenter le modèle théorique et l'équation ajustée.

Le modèle de régression linéaire multiple vise à expliquer la variable dépendante, le score de vengeance au temps 1 (revenge_t1), en fonction de toutes les variables explicatives mesurées au temps 1 : âge, sexe, vc (vindictive complaining), et wom (negative word-of-mouth).

Le modèle théorique va prendre cette forme :

revenge_t1 = β0 + β1 * age + β2 * sexe + β3 * vc + β4 × wom+ ϵ

À partir des résultats de la régression ci-bas, voici l'équation ajustée avec les bons coefficients :

revenge_t1 = −1.26802 + 0.04693 * age + 0.48555 * sexe + 0.57119 * vc + 0.43664 * wom + ϵ

Nb d'observations lues	80
Nb d'obs. utilisées	80

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	4	117.11036	29.27759	102.60	<.0001
Erreur	75	21.40164	0.28536		
Total sommes corrigées	79	138.51200			

Root MSE	0.53419	R carré	0.8455
Moyenne dépendante	4.56000	R car. ajust.	0.8372
Coeff Var	11.71462		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	-1.26802	0.41756	-3.04	0.0033
sexe	1	0.48555	0.13091	3.71	0.0004
age	1	0.04693	0.00874	5.37	<.0001
vc	1	0.57119	0.03785	15.09	<.0001
wom	1	0.43664	0.04800	9.10	<.0001

b. Discuter de la significativité globale du modèle

La valeur F des résultats permet d'effectuer un test d'hypothèse global qui évalue si au moins une des variables explicatives (sexe, âge, vc et wom) a un effet significatif sur la variable dépendante (revenge_t1). L'hypothèse nulle qu'on cherche à prouver est : H0 = β1 = β2 = … = βk = 0 où tous les coefficients de régression sont égaux à zéro. Avec une valeur de 102,60 et une p-value < 0.0001 (inférieure au seuil de 0,05), on rejette l'hypothèse nulle. On conclut donc qu'au moins une des variables explicatives (sexe, âge, vc et wom) a un effet significatif sur la variable dépendante, le score de vengeance au temps 1 (revenge_t1).

Par contre, le test F global ne spécifie pas quelles variables sont significatives. Pour le savoir, il faut examiner les p-value associées à chaque coefficient des variables explicatives.

Variable dépendante : revenge_t1

Nb d'observations lues	80
Nb d'obs. utilisées	80

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	4	117.11036	29.27759	102.60	<.0001
Erreur	75	21.40164	0.28536		
Total sommes corrigées	79	138.51200			

Root MSE	0.53419	R carré	0.8455
Moyenne dépendante	4.56000	R car. ajust.	0.8372
Coeff Var	11.71462		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	-1.26802	0.41756	-3.04	0.0033
age	1	0.04693	0.00874	5.37	<.0001
sexe	1	0.48555	0.13091	3.71	0.0004
vc	1	0.57119	0.03785	15.09	<.0001
wom	1	0.43664	0.04800	9.10	<.0001

On voit que toutes les valeurs des p-values sont bien inférieures à 0,05. On peut donc conclure que toutes les variables explicatives (sexe, âge, vc et wom) sont statistiquement significatives dans le modèle global.

c. Tester et Interpréter l'effet de la variable « vc » sur revenge_t1.

Dans la régression linéaire multiple, le test-t est utilisé pour tester chaque coefficient de régression individuellement. Il permet aussi de tester l'hypothèse nulle afin de déterminer si la variable « vc » a un effet statistiquement significatif sur la variable revenge_t1.

L'hypothèse nulle de ce test est donc : $H0 : \beta_{vc} = 0$

La valeur t obtenu pour vc est 15,09 et donc différente de zéro. Cela signifie que la variable vc a un effet significatif sur revenge_t1 et qu'on rejette l'hypothèse nulle (H0).

De plus, le fait que la p-value est inférieure à 0,0001 indique que cet effet est hautement significatif (<0.05, donc on rejette l'hypothèse nulle). En résumé, vc a un effet significatif et positif sur le score de vengeance.

En analysant les résultats des coefficients estimés, on remarque une valeur de paramètre de 0,57119 pour la variable vc. L'interprétation est donc que, toutes choses étant égales par ailleurs, une augmentation d'une unité de vc est associée à une augmentation moyenne de 0,57119 points de revenge_t1.

d. Tester et Interpréter l'effet de la variable « wom » sur revenge_t1.

On va faire le même exercice effectué à la sous-question c) pour la variable « wom ».

L'hypothèse nulle de ce test sera donc : $H0 : \beta_{wom} = 0$

La valeur t obtenu pour wom est de 9.10 et donc différente de zéro, ce qui signifie que la variable wom a un effet significatif sur revenge_t1 et qu'on rejette l'hypothèse nulle (H0).

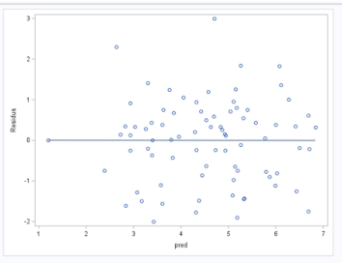
De plus, la p-value est inférieure à 0,0001, donc cet effet est hautement significatif (<0.05) et on rejette l'hypothèse nulle (H0). En résumé, wom a un effet significatif et positif sur le score de vengeance.

En analysant les résultats des coefficients estimés, on remarque une valeur de paramètre de 0,43664 pour la variable wom. L'interprétation est donc que, toutes choses étant égales par ailleurs, une augmentation d'une unité de wom est associée à une augmentation moyenne de 0,43664 points de revenge_t1

e. Vérifier graphiquement les conditions de validité du modèle

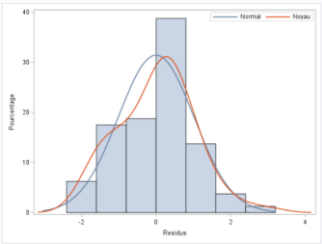
Afin de bien vérifier les conditions de validité du modèle, analysons les graphiques des résidus. Commençons par calculer les résidus et valeurs prédites des RJS avant de créer un graphique des résidus en fonction des valeurs prédites dans le modèle de régression.

```
PROC GLM data=infe.Devoir1;
class sexe ;
model revenge_t1 = age sexe vc wom;
output out=residuals predicted=pred rstudent=Residus;
run;

PROC SGSCATTER data=residuals;
plot Residus*pred / reg=(degree=1);
run;
```



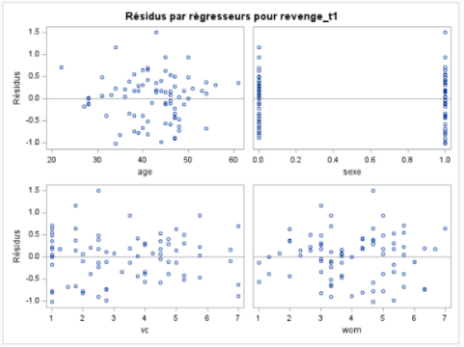
D’abord, dans un modèle de régression bien ajusté, les résidus doivent être répartis aléatoirement autour de zéro, avec des erreurs résiduelles équilibrées entre valeurs positives et négatives à chaque prédiction. Ici, les points sont globalement répartis de manière aléatoire autour de zéro. Cela indique que les résidus ne montrent aucun biais dans les prévisions et que le modèle capture bien la relation entre les variables explicatives et la variable dépendante. Il y a donc un bon équilibre des erreurs résiduelles autour de zéro et aucune preuve claire d'hétéroscédasticité dans la dispersion des points. Ce graphique démontre donc que le modèle respecte globalement les hypothèses fondamentales de linéarité et d'homoscédasticité.

```
PROC GLM data=infe.Devoir1;
class sexe ;
model revenge_t1 = age sexe vc wom;
output out=residuals predicted=pred rstudent=Residus;
run;

PROC SGPLOT data=residuals;
histogram Residus;
density Residus;
density Residus / type=kernel;
keylegend / location=inside position=topright;
run;
```



Ensuite, vérifions si la distribution des résidus suit une distribution normale à l’aide de cet histogramme. Ce graphique permet de vérifier l’hypothèse de normalité des résidus dans le modèle de régression.

L'histogramme démontre que la plupart des résidus sont proches de zéro, avec une répartition qui semble suivre, dans une certaine mesure, une courbe en cloche. Cela indique que les erreurs sont globalement normales, bien que l'on observe une légère asymétrie vers la droite. En effet, quelques résidus sont plus élevés que prévu avec une grande concentration de résidus autour de 0,2. Toutefois, cet écart n'est pas dramatique puisqu’il y a tout de même plusieurs résidus négatifs. Cette petite déviation de la normalité ne représente donc pas un problème majeur pour la validité du modèle. On remarque aussi quelques valeurs extrêmes présentes dans les queues de la distribution, à gauche et à droite. Cependant, ces valeurs ne semblent pas assez extrêmes au point d’invalider directement le modèle.



Terminons en analysant les graphiques de relation entre les prédicteurs (variables explicatives) et les résidus pour détecter s’il y a des relations non linéaires ou de l’hétéroscédasticité entre certaines variables.

- **Âge** : Dans le graphique des résidus par rapport à l'âge, les résidus semblent assez bien répartis aléatoirement autour de zéro. Il n'y a donc pas de signe clair de non-linéarité pour cette variable.
- **Sexe** : Puisque la variable sexe est binaire, les résidus sont empilés verticalement autour de 0 et 1. Il n'y a pas de tendance visible et les résidus sont bien répartis autour de zéro. Cela confirme que l'effet du sexe semble correctement modélisé dans le cadre d'une relation linéaire.
- **VC et WOM** : Pour les variables vc et wom, les résidus semblent être globalement bien répartis aléatoirement autour de zéro. Toutefois, on remarque une légère concentration de résidus pour les points proches de la valeur 1, sur l'axe des X, de la variable vc. Malgré ce léger regroupement de données pour vc, les deux variables peuvent toujours être considérées comme valides dans le modèle de régression, car il n'y a pas de signe d'hétéroscédasticité ou de non-linéarité majeure.

Dans l'ensemble, les diagnostics des graphiques confirment que le modèle de régression multiple est valide. Les résidus sont correctement répartis de manière aléatoire autour de zéro, la normalité est globalement respectée, et il n'y a pas de preuve majeure de non-linéarité ou d'hétéroscédasticité problématique.

f. Laquelle des variables incluses dans le modèle impacte le plus le score de vengeance au temps 1 ? Justifiez adéquatement votre réponse

Pour répondre à cette sous-question, il faut examiner les coefficients estimés dans le modèle de régression ainsi que leurs valeurs de test t afin de déterminer laquelle des variables explicatives possède l'impact le plus important sur le score de vengeance (revenge_t1) au temps 1.

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	-1.26802	0.41756	-3.04	0.0033
sexe	1	0.48555	0.13091	3.71	0.0004
age	1	0.04693	0.00874	5.37	<.0001
vc	1	0.57119	0.03785	15.09	<.0001
wom	1	0.43664	0.04800	9.10	<.0001

On remarque que vc a le plus grand coefficient avec une valeur de 0,57119, ce qui indique que c'est la variable ayant le plus grand effet sur revenge_t1. On peut interpréter le tout de cette façon :

Pour chaque unité d'augmentation de VC, le score de vengeance (revenge_t1) augmente en moyenne de 0.57119 points, toutes choses étant égales par ailleurs.

De plus, vc possède également la valeur de test-t le plus élevé (15.09), ce qui souligne que cet effet est hautement significatif et que « vc » a une influence dominante sur le modèle.

Pour conclure, en tenant compte des réponses données aux sous-questions a) à f), on peut bel et bien affirmer que le sexe a un effet significatif sur le désir de vengeance au temps 1, même lorsque l'on tient compte des autres variables explicatives dans le modèle (âge, vc, et wom). Cet effet est positif et statistiquement significatif tel que prouvé précedemment. Les femmes (sexe = 1) ont un score de vengeance plus élevé que les hommes (sexe = 0), avec une augmentation moyenne de 0.48555 points sur le score de vengeance, toutes choses étant égales par ailleurs. Cette conclusion est soutenue par la p-value faible et la valeur du test-t élevée, qui indiquent que cet effet est très peu susceptible d'être dû au hasard.