

Étude économétrique

Quels sont les déterminants de la baisse des naissances en France depuis 2010 ?

1991 - 2019

*Andrea Chahwan
Lauriane Delpont
Mehdi Fehri
Niclette Ndaya Nsabua*



SOMMAIRE

1

Introduction

Contexte, problématique de l'étude et choix des variables

2

Choix des données et préparation

Collecte des données, définition des variables et période d'analyse

3

Analyse économétrique avec R

Préparation des données, détection et gestion des problèmes spécifiques et choix des modèles

4

Résultats et interpretations

Présentation des principaux résultats et analyse critique

5

Conclusion

Introduction

A. Choix de la variable dépendante

- **678 000 naissances en 2023, niveau le plus bas depuis 1938 (hors-guerre)**
- "Réarmement démographique" (E. Macron)
- Notre étude : **Analyse du taux de fécondité pour comprendre les déterminants de cette baisse.**
- Période couverte : **1991 - 2019**
- **Variable dépendante :**
 - **Nombre de naissance ?**
 - **Taux de natalité ou taux de fécondité ?**



Introduction

B.Les variables explicatives

Facteurs considérés pour le choix des variables explicatives :

- Socio-culturels
- Démographiques et biologiques
- Économiques et politiques
- Travail et éducation
- Individuels et comportementaux

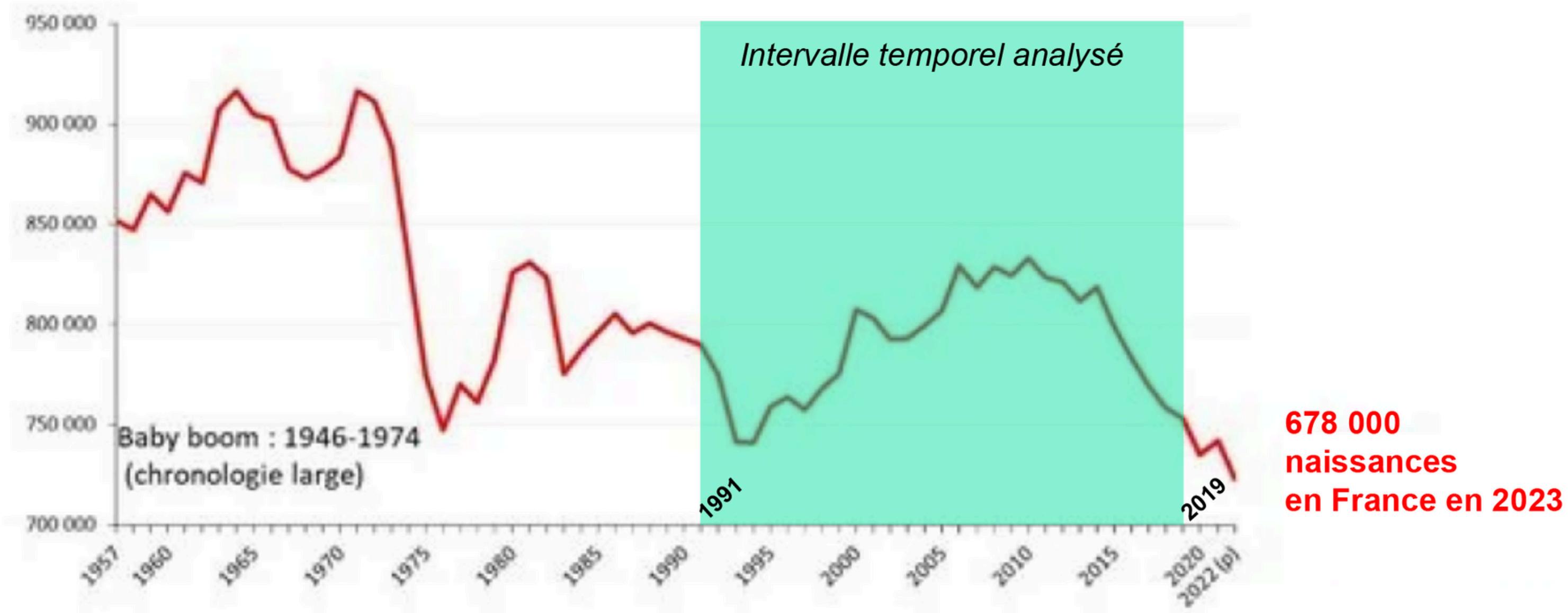


Nous avons en tout sélectionnés 37 données

Introduction : Intervalle périodique analysé



Le nombre de naissance atteint un point bas historique



Champ : France hors Mayotte jusqu'en 2013, y compris Mayotte à partir de 2014.

Source : Insee, statistiques de l'état civil (résultats 2022 provisoires arrêtés à fin novembre 2022)

Introduction

Les sources de nos données



eurostat



Premier traitement des variables explicatives

Dictionnaire des variables explicatives

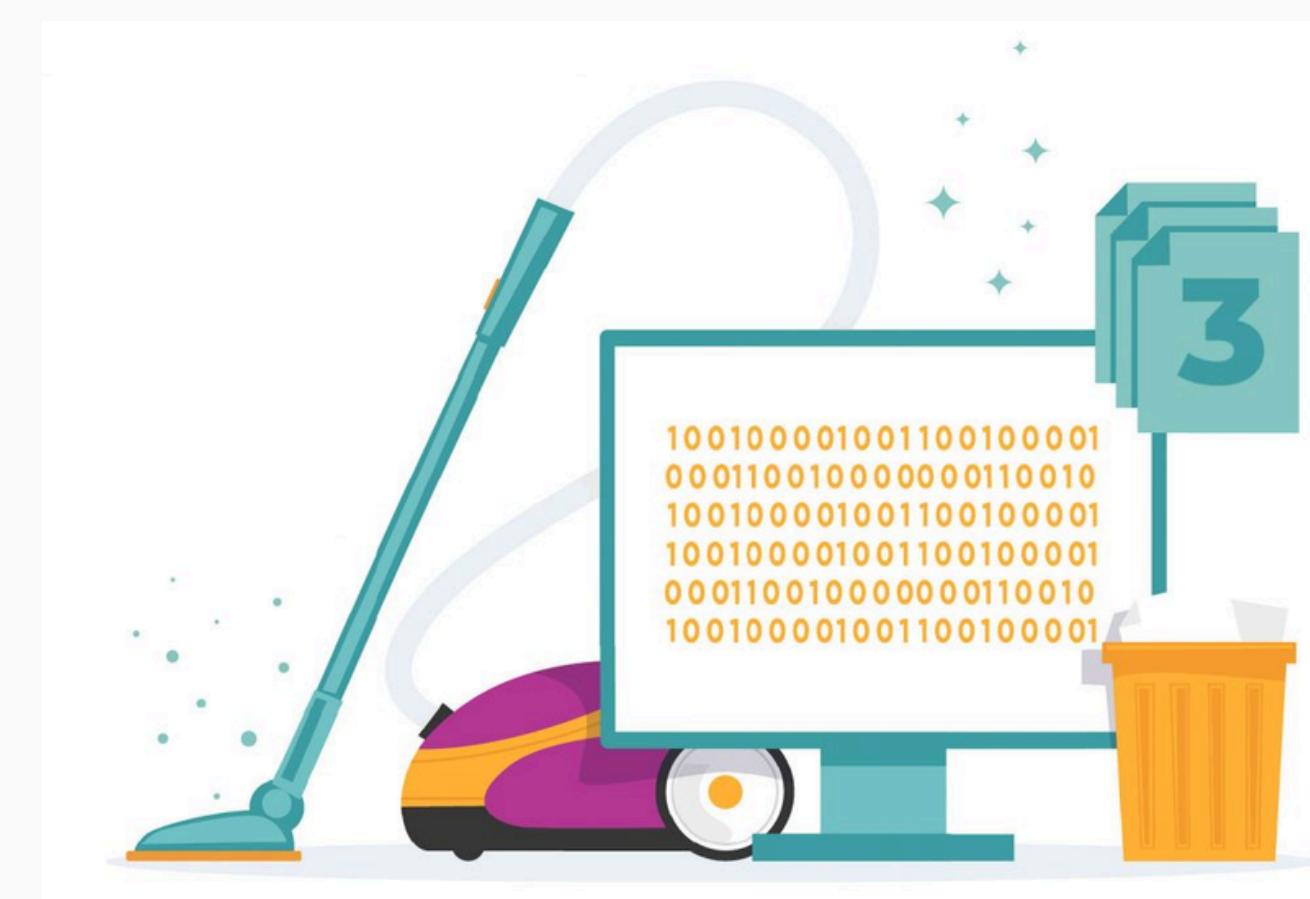
Variable	Definition
pa_synthé	Pouvoir d'achat synthétique
preca_fem	Taux de précarité de l'emploi des femmes
tpspartiel	Quotité de temps de travail partiel
bourse	Cours des actions
deficit	Déficit de l'État
opi_ameli_niv_vie	A
opi_inflation	Opinion sur l'évolution future des prix
opi_env	Préoccupation pour l'environnement
nuptialite	Taux de nuptialité
spepro_cadrefemmes	Spécialisation professionnelle des femmes
ivg_100	Nombre d'IVG pour 100 naissances
étude_sup	Proportion de la population ayant poursuivi des études supérieures au bac

Variables suspectées d'être endogènes

Choix des données et préparation

Nettoyage en amont des données

- Synthétisation de certaines données
- Gestion des données manquantes
- Ajustement pour uniformiser les variables
- Sélection des variables les plus pertinentes et les moins redondantes



Les enseignements qui ont guidé la construction de notre code final

- Lors de nos premiers run et tests de code, nous avons identifié deux problèmes majeurs et persistants susceptibles de complexifier notre travail :
 1. **Les problèmes d'hétéroscédasticité**
 2. **Les problèmes d'autocorrélation** importante et à plusieurs niveaux (lags) (logique quand on traite des séries chronologiques)

**Tout le code que nous avons construit est axé autour de l'objectif de réduire au maximum ces deux problématiques (tout en traitant les autres spécificités)*

Analyse économétrique avec R

I - Préparation des données pour le modèle économétrique

1. Nettoyage des données et sélection des variables (39 variables -> 15)
2. Visualisation initiale et graphique des variables explicatives
3. **Interpolation trimestrielle: (29 observations -> 109)**
 - i. Plus de degré de liberté
 - ii. Meilleure précision des estimations ($N \rightarrow +\infty$)
 - iii. Tests statistiques plus fiables
4. **Création de variables retardées (lags) : les conditions de l'année N influencent les naissances de N+1**
 - i. Lien temporel réaliste : 9 mois de gestation avant naissance

Analyse économétrique avec R

I - Préparation des données pour le modèle économétrique

1. Transformation de notre série chronologique en série (détrendisation + différenciation

stationnaire, pourquoi ? :

- a. **Stabilité des propriétés statistiques** (variance, moyenne et covariance constantes) --> Homoscédasticité
- b. **Réduire le biais lié à l'évolution temporelle** (on veut des variations indépendantes dans le temps) pour une estimative qui reflète les vraies dynamiques sous-jacentes.
- c. **Ciblage autour des fluctuations à court terme** : Plus pertinente
- d. **Condition nécessaire pour traiter l'autocorrélation et d'autres problématiques.** Les solutions qui exigent une condition de stationnarité (GLS, ARIMA, Test DW, BP, solutions robustes Newey-West etc..)

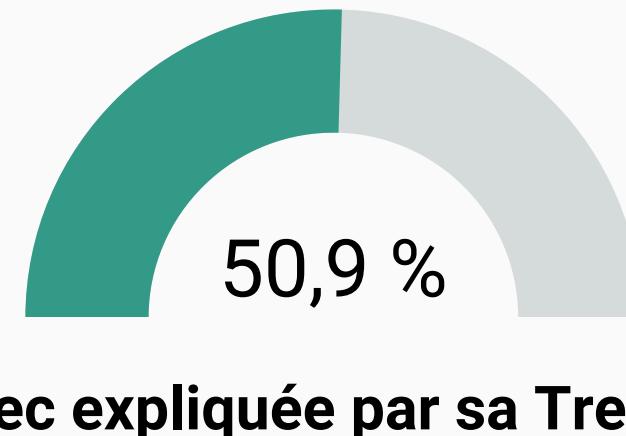
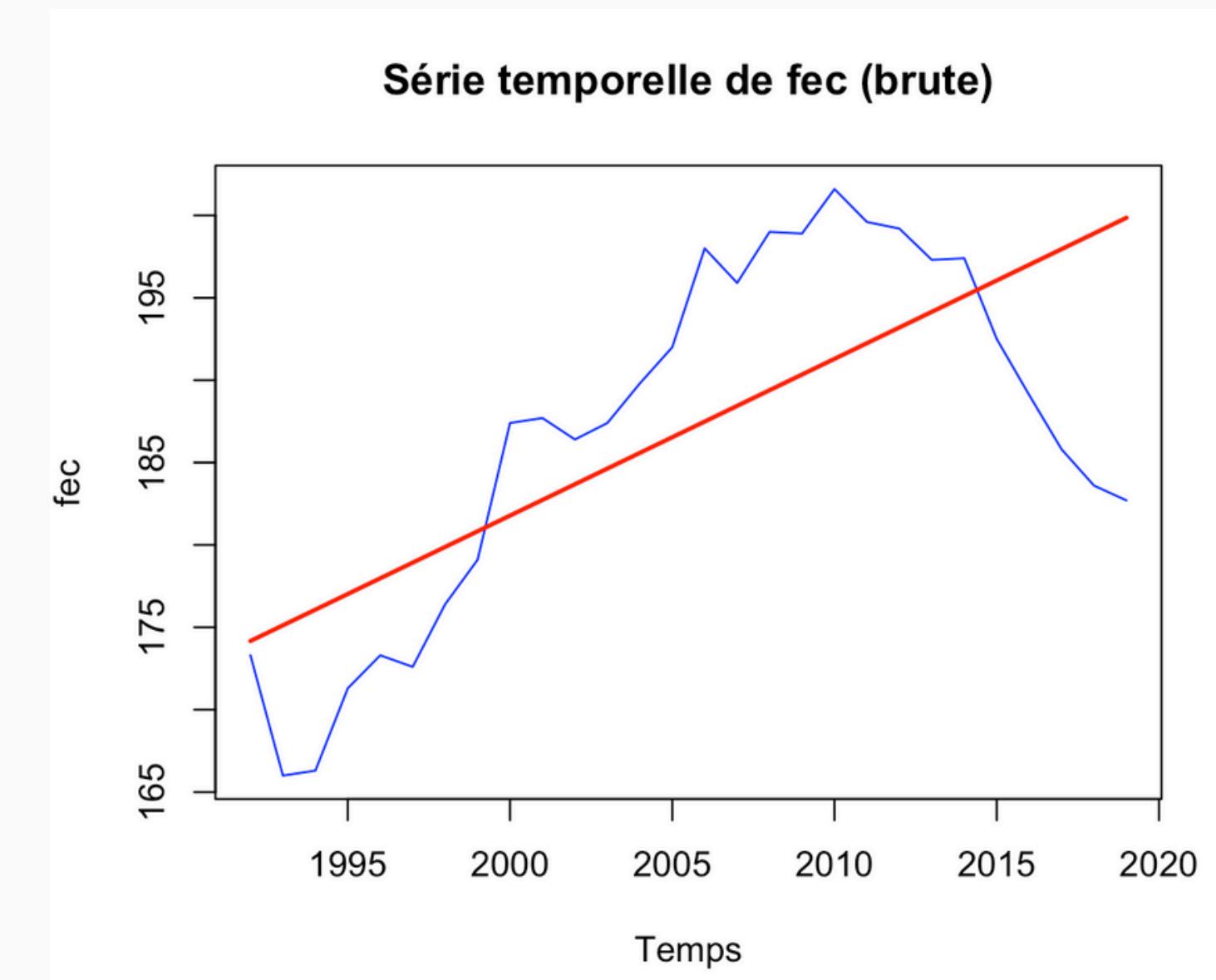
Analyse économétrique avec R

I - Préparation des données pour le modèle économétrique

- Enseignement après détrendisation et différenciation de la série

1. Test ADF confirme la stationnarité après la différenciation.
- 2. Part importante de l'évolution du taux de fécondité est expliquée par la tendance**

-> Transformation de notre variable dépendante en variable différenciées (variation absolue) que nous avons transformée fine en variation relative = taux de croissance périodique



En transformant notre variable d'intérêt en taux de variation, on répond à une nouvelle question :

Quels facteurs influencent les fluctuations du taux de fécondité d'une période à l'autre ?

-> L'analyse change et devient mécaniquement une semi-élasticité (variation en % de y)

Analyse économétrique avec R

I - Préparation des données pour le modèle économétrique, la suite...

1. Calculs de R^2 individualisé de chaque variable explicative
2. **Standardisation du modèle avant** traitement des outliers:
 - i. **échelle homogène** pour mieux comparer les variables (réduit les biais des fortes amplitudes).
 - ii. rend la détection des écarts plus fiable.
3. **Traitement des outliers** et création d'un nouveau jeu de données sans outliers
 - i. **Avantage** : améliore notre R^2 , facilite l'interprétation, permet d'assurer que les résidus suivent une distribution normalisée, réduit l'influence des valeurs extrêmes, homogénise la variance.

Résidus Standardisés (Zoomé)

Résidus standardisés

1

0

-1

-2

0 30 60 90
Index d'observation

Suppression des outliers au delà d'un écart-type de 2 (modèle standardisé)

- **Inconvénients :** suppressions de 4 observations

Suppression des variables avec un VIF élevé et avec des colinéarités parfaites

Résultats de la régression OLS avant traitement des outliers

```

Residuals:
    Min      1Q   Median      3Q     Max 
-0.0061270 -0.0019389  0.0001054  0.0019985  0.0051674 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         0.0004785  0.0002633  1.818  0.072321 .  
lag_pa_synthé                      0.0258654  0.0084370  3.066  0.002835 ** 
lag_preca_fem                     -0.0042848  0.0024749 -1.731  0.086684 .  
lag_tpspartiel                     0.0099723  0.0015533  6.420  5.51e-09 *** 
lag_bourse                          -0.0010478  0.0012728 -0.823  0.412434  
lag_deficit                         -0.0003129  0.0011273 -0.278  0.781953  
lag_opi_inflation                   -0.0053902  0.0043326 -1.244  0.216555  
lag_opi_amelio_niv_vie            -0.0028451  0.0011948 -2.381  0.019274 *  
lag_opi_env                          0.0009002  0.0010700  0.841  0.402314  
lag_nuptialite                     -0.0002351  0.0017382 -0.135  0.892688  
lag_naisshorsma                    -0.0420363  0.0111368 -3.775  0.000281 *** 
lag_spepro_cadrefemmes             0.0055794  0.0043060  1.296  0.198242  
lag_ivg_100                          0.0019572  0.0007196  2.720  0.007781 ** 
lag_étude_sup                       0.0126776  0.0083960  1.510  0.134410  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002736 on 94 degrees of freedom
Multiple R-squared:  0.6877,    Adjusted R-squared:  0.6445 
F-statistic: 15.92 on 13 and 94 DF,  p-value: < 2.2e-16

```

Identification des paires de variables les plus corrélées

```
print(high_corr_pairs)
```

	Var1	Var2	value
0	lag_spepro_cadrefemmes	lag_naisshorsma	0.9919051
1	lag_spepro_cadrefemmes	lag_pa_synthé	0.9912237
2	lag_naisshorsma	lag_pa_synthé	0.9875678
3	lag_étude_sup	lag_naisshorsma	0.9815468
4	lag_étude_sup	lag_spepro_cadrefemmes	0.9712397
5	lag_étude_sup	lag_opi_inflation	0.9706926
6	lag_étude_sup	lag_pa_synthé	0.9615691
7	lag_naisshorsma	lag_opi_inflation	0.9239915
8	lag_étude_sup	lag_preca_fem	0.9188064
9	lag_naisshorsma	lag_preca_fem	0.9161032
10	lag_spepro_cadrefemmes	lag_opi_inflation	0.9136775

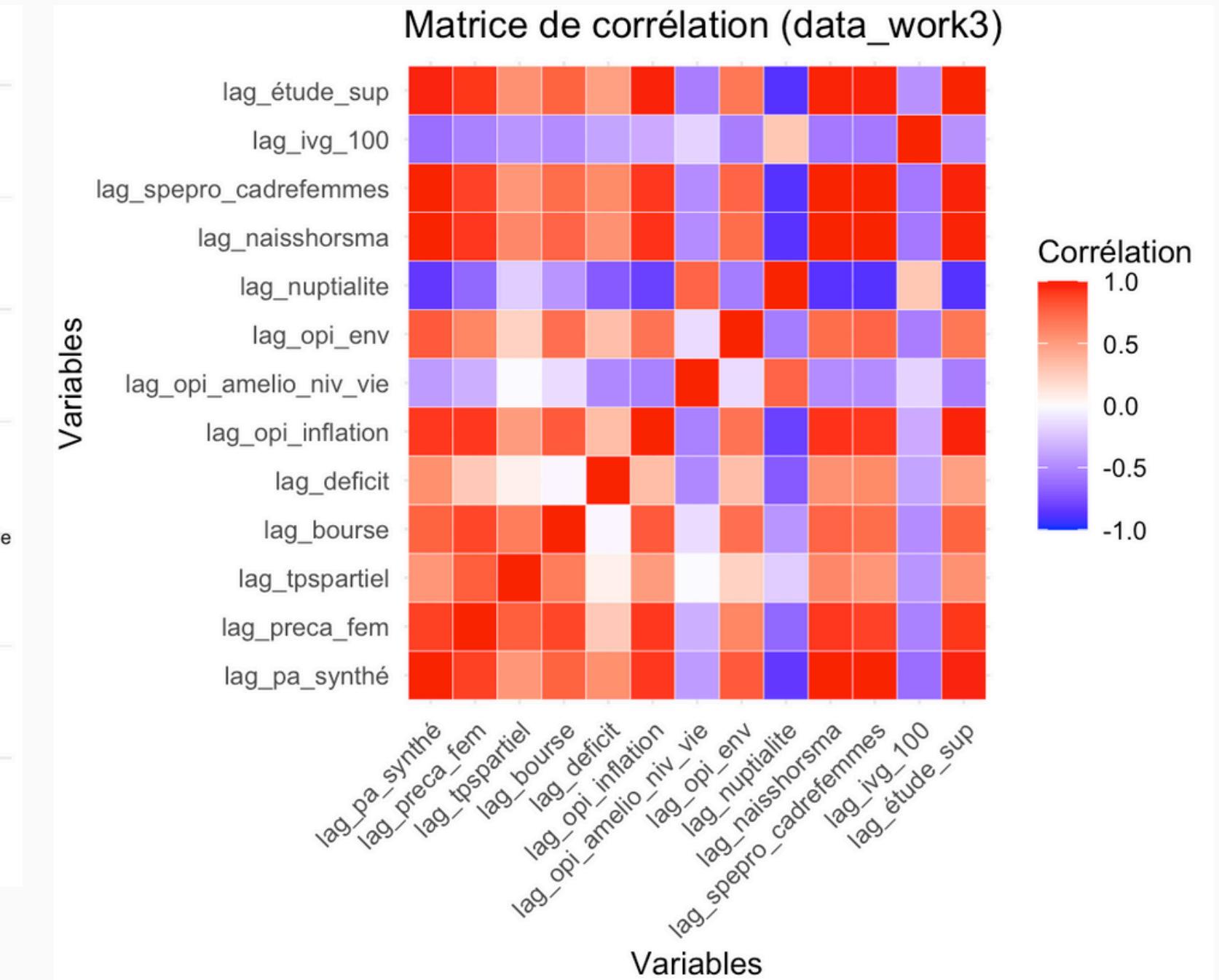
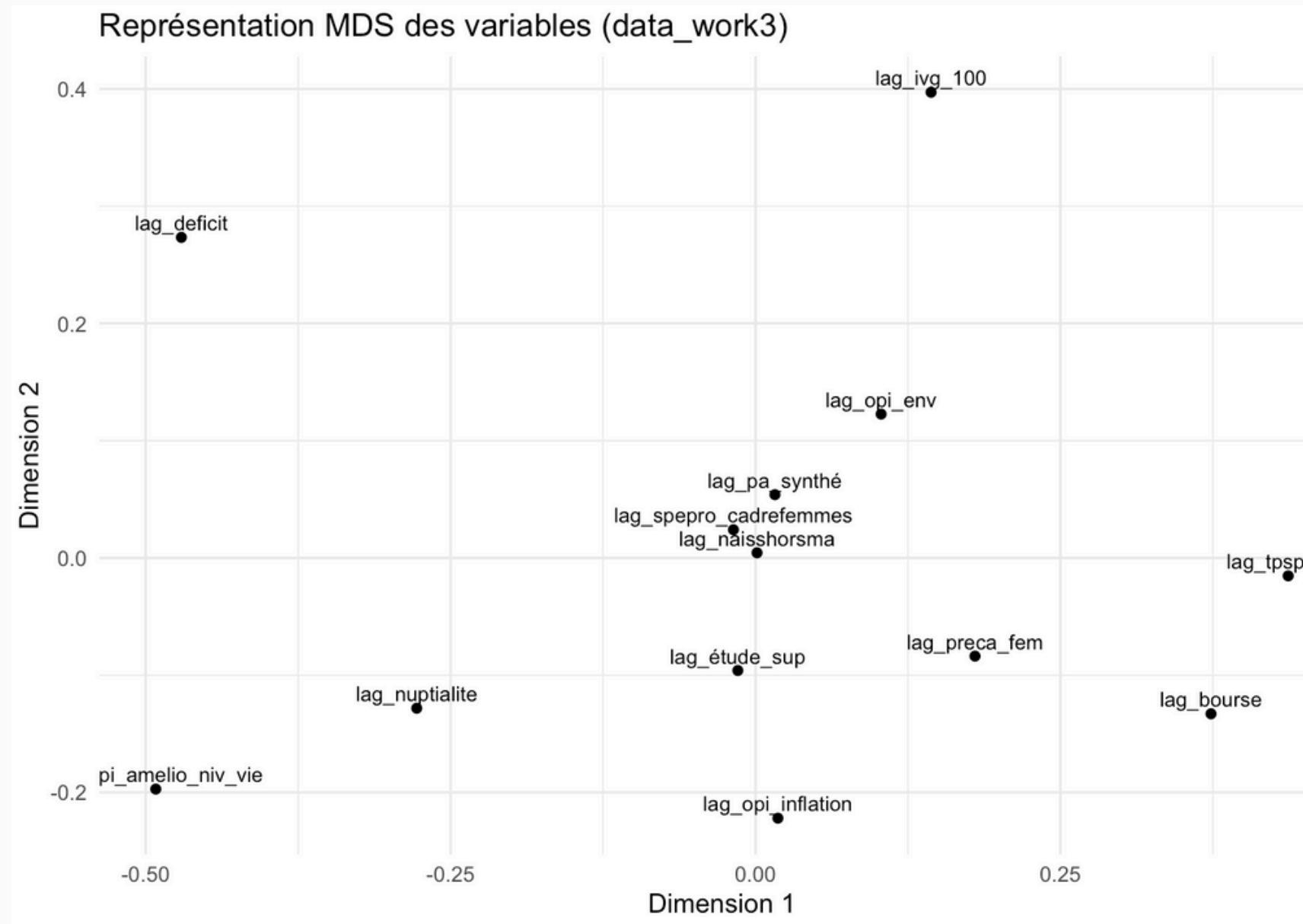
```
### Variables les plus corrélées ###
> print(variable_counts)
# A tibble: 12 × 2
  Variable           Frequency
  <fct>              <int>
1 lag_pa_synthé          9
2 lag_naisshorsma         9
3 lag_spapro_cadrefemmes  9
4 lag_preca_fem            8
5 lag_bourse                8
6 lag_opi_inflation         8
7 lag_étude_sup               8
8 lag_tpspartiel             7
9 lag_opi_env                  7
10 lag_deficit                 3
11 lag_opi_amelio_niv_vie        1
12 lag_nuptialite                 1
```

Analyse économétrique avec R

II – Traitement pour réduire les risques d'erreur de spécification, de biais et d'instabilité du modèle, et à améliorer la pertinence, la robustesse et la précision des estimations.

1. Détection et traitement de la colinéarité parfaite (alias)
2. Analyse de corrélation entre variables explicatives (matrice de corrélation, heatmap, MDS)
3. Réduction des variables à fort VIF (Facteur d'inflation de la variance) et création du DataFrame final

C'est sur cette base que nous avons sélectionné les variables les plus pertinentes que nous voulions garder



--- Itération 1 ---

Facteurs d'inflation de la variance (VIF) actuels :

<code>lag_pa_synthé</code>	<code>lag_preca_fem</code>
1063.723204	94.986305
<code>lag_tpsspartiel</code>	<code>lag_bourse</code>
35.601785	23.941494
<code>lag_deficit</code>	<code>lag_opi_inflation</code>
19.104356	275.066888
<code>lag_opi_amelio_niv_vie</code>	<code>lag_opi_env</code>
21.776302	16.283717
<code>lag_nuptialite</code>	<code>lag_naisshorsma</code>
44.805524	1807.325333
<code>lag_spepro_cadrefemmes</code>	<code>lag_ivg_100</code>
288.188246	7.841978
<code>lag_étude_sup</code>	
986.196344	

Variable avec le VIF le plus élevé : `lag_naisshorsma` (1807.325)

--- Itération 2 ---

Facteurs d'inflation de la variance (VIF) actuels :

<code>lag_pa_synthé</code>	<code>lag_preca_fem</code>
546.510078	94.857912
<code>lag_tpsspartiel</code>	<code>lag_bourse</code>
17.669629	22.788419
<code>lag_deficit</code>	<code>lag_opi_inflation</code>
17.672212	118.998591
<code>lag_opi_amelio_niv_vie</code>	<code>lag_opi_env</code>
5.983216	14.539548
<code>lag_nuptialite</code>	<code>lag_spepro_cadrefemmes</code>
43.410577	282.806049
<code>lag_ivg_100</code>	<code>lag_étude_sup</code>
7.571946	264.207775

Toutes les variables ont un VIF <= 666 . Fin de la boucle.

-> Analyse itérative des VIF et suppression automatique des variables avec les VIF au delà du seuil que nous avons fixé :

- Exemple : suppression de la variable “naissance hors mariage”.

Analyse économétrique avec R

II – Traitement pour réduire les risques d'erreur de spécification, de biais et d'instabilité du modèle, et à améliorer la pertinence, la robustesse et la précision des estimations.

1. Tests de spécification (RESET), choix de la forme fonctionnelle
2. Choix de la meilleure transformation : transformation logarithmique des variables explicatives et régression OLS (interprétation double log ?)

	Model	Adjusted_R2	AIC	BIC	RESET_p_value	RESET_Interpretation
1	Model OLS	0.6024655	-915.6126	-878.5912	0.09429862	Correctement spécifié
2	Model OLS (Log)	0.6446808	-927.2882	-890.2668	0.23075053	Correctement spécifié

-> Choix de transformer nos variables explicatives en log pour interprétation finale équivalente à un double log : ÉLASTICITÉ

Residuals:

	Min	1Q	Median	3Q	Max
	-0.0058845	-0.0011722	0.0001364	0.0013188	0.0065064

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.216308	0.227131	0.952	0.343559
log_lag_pa_synthé	-0.174425	0.038946	-4.479	2.27e-05 ***
log_lag_preca_fem	-0.013708	0.019677	-0.697	0.487867
log_lag_tpspartiel	0.060950	0.021308	2.860	0.005296 **
log_lag_bourse	0.019317	0.004193	4.607	1.39e-05 ***
log_lag_deficit	0.013926	0.002242	6.211	1.75e-08 ***
log_lag_opi_inflation	0.066211	0.025657	2.581	0.011538 *
log_lag_opi_amelio_niv_vie	0.014455	0.003696	3.911	0.000182 ***
log_lag_opi_env	0.007718	0.003100	2.490	0.014671 *
log_lag_nuptialite	-0.007631	0.012115	-0.630	0.530444
log_lag_spepro_cadrefemmes	0.061821	0.015511	3.986	0.000140 ***
log_lag_ivg_100	0.032391	0.020110	1.611	0.110872
log_lag étude_sup	-0.036704	0.016885	-2.174	0.032436 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002163 on 87 degrees of freedom

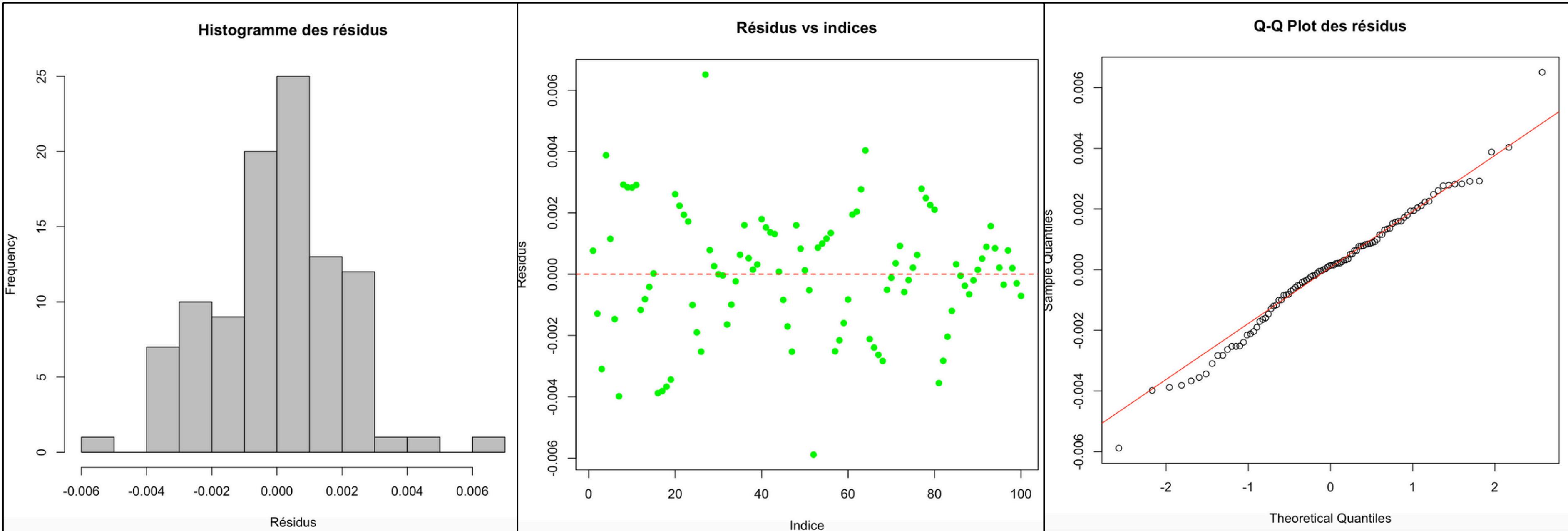
Multiple R-squared: 0.7653, Adjusted R-squared: 0.7329

F-statistic: 23.64 on 12 and 87 DF, p-value: < 2.2e-16

Modèle finale :
100 observations et 12 variables explicatives

Analyse économétrique avec R

VISUALISATION DES RESIDUS DE NOTRE MODELE FINAL



Analyse économétrique avec R

III – Les 4 + 1 Hypothèses de Gauss-Markov : Diagnostic et Traitement des problématiques

Préparation des données instrumentales pour traiter l'endogénéité :

Hypothèse 1 : Moyenne et Somme des résidus doit être nulle

Hypothèse 2 : Endogénéité : Analyse des risques d'endogénéité, identification de variables endogènes
identification des variables candidate à l'instrumentalisation Test de la pertinence des instruments (First stage), estimation IV, Test de Sargan, Test de Hausman

Hypothèse 3 : Homoscédasticité : Analyse et visualisation de la forme d'hétéroscédasticité, Tests d'hétéroscédasticité (Breusch-Pagan, White)

Hypothèse 4 : Analyse de la forme d'autocorélation, détection de stationnarité ou de saisonnalité de la série, tests d'autocorrélation : (Durbin-Watson, Breusch-Godfrey, Ljung-Box). Ajustement d'un modèle ARIMA sur les résidus pour corriger l'autocorrélation, nouvelle régression et vérifications

Conclusions

Vérification finale des hypothèses, construction du modèle final et comparaison des résultats

Analyse économétrique avec R

Normalité des résidus

Test de Shapiro-Wilk :

```
> cat("Statistique W :", round(shapiro_test$statistic, 4), "\n")
Statistique W : 0.9881
> cat("P-value :", round(shapiro_test$p.value, 4), "\n")
P-value : 0.5175
> if (shapiro_test$p.value > 0.05) {
+   cat("Conclusion : Les résidus suivent une distribution normale (H0 acceptée).\n")
+ } else {
+   cat("Conclusion : Les résidus ne suivent pas une distribution normale (H0 rejetée).\n")
+ }
Conclusion : Les résidus suivent une distribution normale (H0 acceptée).
```

Analyse économétrique avec R

Hypothèse 1: Moyenne et Somme des résidus doit être nulle

```
--- Étape 1 : Résidus de moyenne nulle ---
> mean_residuals <- mean(residus)
> std_error_residuals <- sd(residus) / sqrt(length(residus))
> t_stat <- mean_residuals / std_error_residuals
> p_value <- 2 * pt(-abs(t_stat), df = length(residus) - 1)
> cat("Moyenne des résidus :", round(mean_residuals, 5), "\n")
Moyenne des résidus : 0
> cat("t-statistic :", round(t_stat, 3), " | p-value :", round(p_value, 5), "\n")
t-statistic : 0 | p-value : 1
> if (p_value > 0.05) {
+   cat("H0 est vérifiée : les résidus ont une moyenne nulle.\n")
+ } else {
+   cat("H0 est rejetée : les résidus n'ont pas une moyenne nulle.\n")
+ }
H0 est vérifiée : les résidus ont une moyenne nulle.
```

Hypothèse 2: Enxogénéité des variables explicatives

Étape 1: identification de potentiels variables endogènes ?

Variables potentiellement endogènes (significatives au moins une fois) :

```
> print(potentially_endogenous)
      Variable
1   log_lag_pa_synthé
2   log_lag_preca_fem
3   log_lag_tpspartiel
4   log_lag_bourse
5   log_lag_deficit
6 log_lag_opi_amelio_niv_vie
7   log_lag_opi_env
8   log_lag_étude_sup
> cat("\n### Analyse des résidus croisés ###\n")

      Significant
log_lag_pa_synthé.cor    FALSE
log_lag_preca_fem.cor    FALSE
log_lag_tpspartiel.cor   FALSE
log_lag_bourse.cor       FALSE
log_lag_deficit.cor     FALSE
log_lag_opi_inflation.cor FALSE
log_lag_opi_amelio_niv_vie.cor FALSE
log_lag_opi_env.cor     FALSE
log_lag_nuptialite.cor   FALSE
log_lag_spepro_cadrefemmes.cor FALSE
log_lag_ivg_100.cor      FALSE
log_lag_étude_sup.cor    FALSE
> if (any(corr_results$Significant)) {
+   cat("\nCertaines variables explicatives sont significativement corrélées aux résidus. Risque d'endogén")
+ } else {
+   cat("\nAucune variable explicative n'est significativement corrélée aux résidus. Pas d'évidence d'endo")
+ }
```

Aucune variable explicative n'est significativement corrélée aux résidus. Pas d'évidence d'endogénéité.

Test de corrélation entre les variables explicative et les résidus (à un seuil de confiance de 5%)

Hypothèse 2: Enxogénéité des variables explicatives

Étape 1: identification de potentiels variables endogènes ?

```
### Analyse des résidus croisés ###
> # Régression des variables explicatives sur les résidus
> cross_results <- lapply(variables_expliquantes_ols, function(var) {
+   model_residu <- lm(data_work4_log_out[[var]] ~ residus)
+   summary_model <- summary(model_residu)
+   data.frame(
+     Variable = var,
+     Coefficient = coef(summary_model)[["residus", "Estimate"]],
+     P_Value = coef(summary_model)[["residus", "Pr(>|t|)"]],
+     Significant = coef(summary_model)[["residus", "Pr(>|t|)"]] <= 0.05
+   )
+ }) %>% bind_rows()
> print(cross_results)
```

	Variable	Coefficient	P_Value	Significant
1	log_lag_pa_synthé	1.375754e-15	1	FALSE
2	log_lag_preca_fem	2.751509e-15	1	FALSE
3	log_lag_tpspartiel	-4.127263e-15	1	FALSE
4	log_lag_bourse	5.503018e-15	1	FALSE
5	log_lag_deficit	-5.503018e-15	1	FALSE
6	log_lag_opi_inflation	1.375754e-15	1	FALSE
7	log_lag_opi_amelio_niv_vie	-2.063632e-15	1	FALSE
8	log_lag_opi_env	1.582118e-14	1	FALSE
9	log_lag_nuptialite	-2.063632e-15	1	FALSE
10	log_lag_spepro_cadrefemmes	-1.100604e-14	1	FALSE
11	log_lag_ivg_100	-8.598465e-16	1	FALSE
12	log_lag_étude_sup	1.100604e-14	1	FALSE

Analyse des résidus croisés

Traitemet des variables endogènes

Dictionnaire des variables explicatives

Variable	Definition
pa_synthé	Pouvoir d'achat synthétique
preca_fem	Taux de précarité de l'emploi des femmes
tpspartiel	Quotité de temps de travail partiel
bourse	Cours des actions
deficit	Déficit de l'État
opi_ameli_niv_vie	A
opi_inflation	Opinion sur l'évolution future des prix
opi_env	Préoccupation pour l'environnement
nuptialite	Taux de nuptialité
spepro_cadrefemmes	Spécialisation professionnelle des femmes
ivg_100	Nombre d'IVG pour 100 naissances
étude_sup	Proportion de la population ayant poursuivi des études supérieures au bac

Variables suspectées d'être endogènes

Les Variables instrumentales

Dictionnaire des variables instrumentales

Variables Instrumentales	Definition
depseniors	Taux de dépendance des seniors
synthé_Oiseau	Indicateur synthétique relatif aux milieux bâti, agricole et forestier de la population d'oiseaux
lt_interest_rate	Taux d'intérêt à long terme
opi_famille	La famille comme lieu de bien-être et de détente
opi_work_fem	Évolution des opinions sur le travail des femmes
pib_hab	PIB par habitant
acceuilenf	Nombre d'établissements et services d'accueil pour jeunes enfants

Hypothèse 2 : Exogénéité

Explication des variables soupçonnées d'endogénéité

- **preca_fem** : Un taux de fécondité élevé peut augmenter la précarité de l'emploi des femmes en limitant leurs opportunités professionnelles.
- **tpspartiel** : Un taux de fécondité élevé pousse souvent les femmes à opter pour des emplois à temps partiel afin de concilier travail et parentalité.
- **opi_env** : Un taux de fécondité élevé peut accroître les préoccupations environnementales des individus concernant l'avenir de leurs enfants.
- **sepro_cadrefemmes** : Un taux de fécondité élevé peut limiter la spécialisation professionnelle des femmes en réduisant le temps ou les opportunités pour une carrière avancée.
- **étude_sup** : Un taux de fécondité élevé peut freiner la poursuite d'études supérieures, notamment chez les femmes, en raison des responsabilités parentales.

Hypothèse 2 : Exogénéité

Nos Variables instrumentales

- **Taux d'intérêt à long terme** : Indicateur indirect de l'état du marché du travail, lié à la confiance économique, bon prédicteur de crises économiques reflétant aussi les conditions de financement (e.g. : peut affecter les choix des individus à poursuivre les études).
- **Taux de dépendance des seniors** : Réflexion des pressions démographiques et financières qui impactent les choix professionnels et de vie des individus.
- **Indicateur synthétique des milieux bâtis, agricoles et forestiers** : Mesure indirecte des préoccupations liées aux préoccupations environnementales (l'état de la biodiversité peut impacter l'opinion des individus sur l'environnement).

Hypothèse 2 : Exogénéité

Variables instrumentales

- **La famille comme lieu de bien-être et de détente** : Indicateur des valeurs sociales affectant les conditions de travail et la répartition des rôles, lié à l'emploi des femmes.
- **Évolution des opinions sur le travail des femmes** : Révèle les changements sociaux qui influencent la spécialisation professionnelle des femmes.
- **PIB par habitant** : Indicateur indirect du niveau de développement, lié à la montée des préoccupations environnementales.
- **Nombre d'établissements et services d'accueil pour jeunes enfants** : Indicateur du soutien aux femmes actives, impactant la précarité de l'emploi.

Hypothèse 2 : Exogénéité

Variables instrumentales : le modèle 2SLS

1^{re} étape : pertinence des variables

F-stat élevé des instruments sur les variables endogènes

--- Résultats agrégés des premières étapes ---			
	Endogenous_Variable	F_statistic	F_p_value
1	log_lag_preca_fem	13.279351	1.874937e-11
2	log_lag_tpsspartiel	29.613850	2.710403e-20
3	log_lag_opi_env	9.078484	2.515120e-08
4	log_lag_nuptialite	28.161980	1.176502e-19
5	log_lag_spepro_cadrefemmes	35.627193	1.006217e-22
	R_squared	Adj_R_squared	
1	0.9819296	0.9789533	
2	0.9619095	0.9556358	
3	0.9662186	0.9606545	
4	0.9907650	0.9892439	
5	0.9983502	0.9980785	

Interprétation du test de première étape :

- Un F-statistic élevé (souvent > 10) indique que les instruments sont pertinents.

	Pr(> t)
(Intercept)	0.95918
log_lag_pa_synthé	0.10662
log_lag_preca_fem	0.49524
log_lag_tpsspartiel	0.21378
log_lag_bourse	0.06502 .
log_lag_deficit	0.00164 **
log_lag_opi_inflation	0.21422
log_lag_opi_amelio_niv_vie	0.00150 **
log_lag_opi_env	0.05924 .
log_lag_nuptialite	0.39673
log_lag_spepro_cadrefemmes	0.40913
log_lag_ivg_100	0.91657
log_lag_étude_sup	0.26140

Signif. codes:	0 *** 0.001 ** 0.01 * 0.05 . 0.1 ' ' 1
Residual standard error: 0.002319 on 87 degrees of freedom	
Multiple R-Squared: 0.7303, Adjusted R-squared: 0.693	
Wald test: 19.03 on 12 and 87 DF, p-value: < 2.2e-16	

Hypothèse 2 : Exogénéité

Variables instrumentales

Test d'Hausman

Un p-value élevé suggère que le biais d'endogénéité n'est pas significatif, l'OLS pourrait être acceptable.

```
Test de Hausman:
```

```
$statistic  
[1] 2.245675
```

```
$p_value  
[1] 0.9995672
```

Interprétation du Test de Hausman:

- Un p-value faible suggère que le modèle OLS est biaisé et que le modèle IV est préférable.
- Un p-value élevé suggère que le biais d'endogénéité n'est pas significatif, l'OLS pourrait être acceptable.

Test Sargan

Une p-value élevé indique que les instruments sont globalement valides.

```
Test de Sargan:
```

```
> print(sargan_test)  
      df1      df2 statistic   p-value  
2.0000000       NA 3.1530838 0.2066886  
> cat("\nInterprétation du Test de Sargan:\n")
```

Interprétation du Test de Sargan:

```
> cat("- Un p-value faible (<0.05) indique que les instruments ne sont pas valides (suridentification non rejetée).\n")  
- Un p-value faible (<0.05) indique que les instruments ne sont pas valides (suridentification non rejetée).
```

Hypothèse 2 : Exogénéité

Conclusion

```
==== Tableau récapitulatif des résultats ===
> print(summary_table)

                                         Test
1 First-Stage Regression (log_lag_confiance_menage)
2 First-Stage Regression (log_lag_emploi_niveau_vie)
3                                         Sargan Test
4                                         Hausman Test

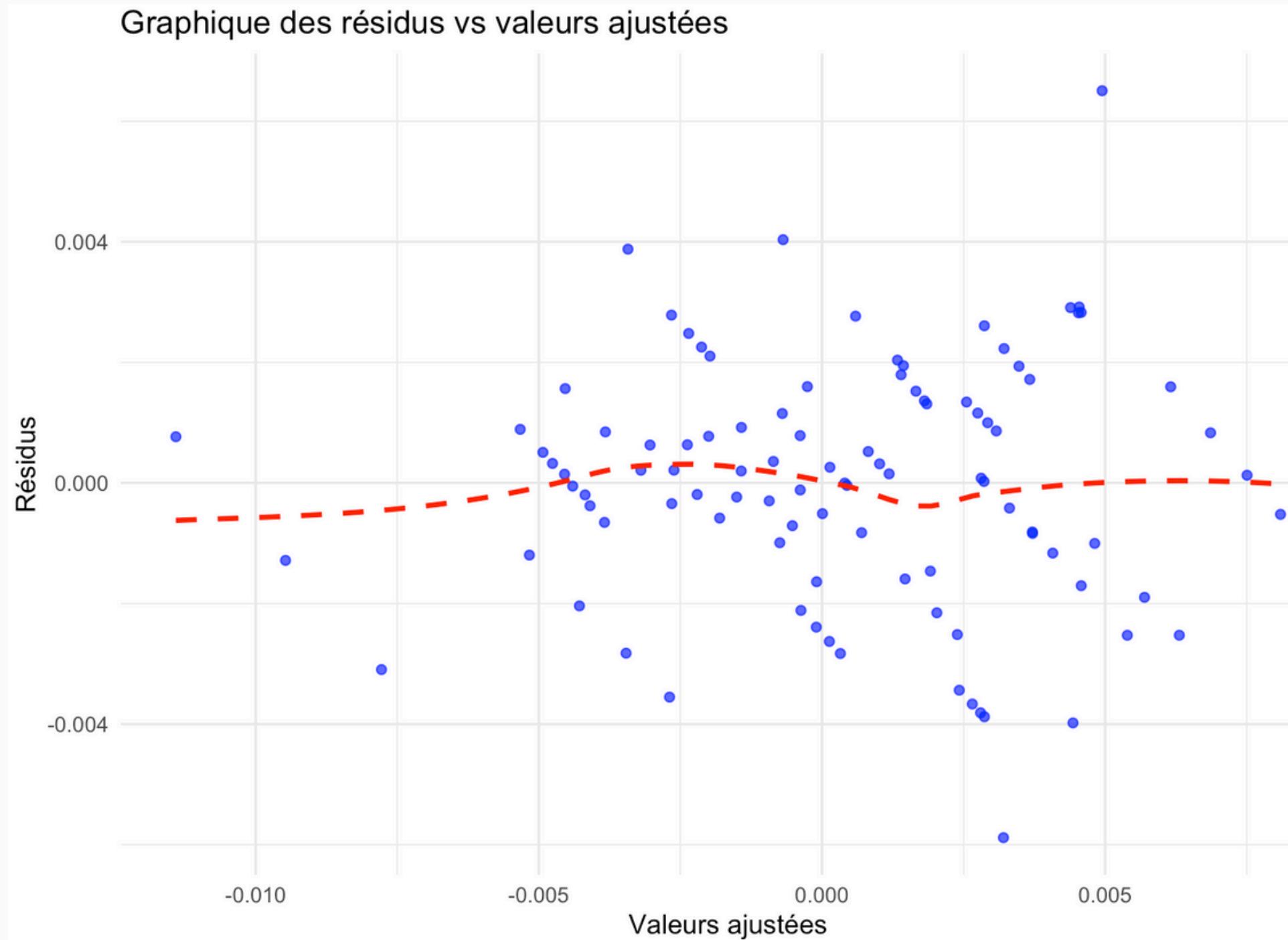
      Statistic      Value      P_Value
1 F-statistic 13.279351 1.874937e-11
2 F-statistic 29.613850 2.710403e-20
3 Chi-Squared  3.153084 2.066886e-01
4 Chi-Squared  2.245675 9.995672e-01

                                         Conclusion
1 Instruments are strong and relevant for log_lag_confiance_menage.
2 Instruments are strong and relevant for log_lag_emploi_niveau_vie.
3                                         Instruments are globally valid.
4 OLS is acceptable as no significant endogeneity bias is detected.
```

Hypothèse 3 : Homoscédasticité

FORME D'UNE PROBABLE HÉTÉRO ?

Graphique des résidus vs valeurs ajustées



- Avant ce code final, on avait un gros problème d'hétérogénéité qu'on n'avait pas du tout réussi à corriger malgré nos tentatives.
- Nous avions envisagé d'utiliser les **SE Robustes de White**.
- Étant donné le second problème d'autocorrélation, nous envisageons peut être aussi d'utiliser des erreurs robustes de **Newey-West** voire de transformer notre modèle en **GLS**.

Test de Breusch-Pagan

Residuals:

	Min	1Q	Median	3Q	Max
	-8.527e-06	-3.294e-06	-8.630e-07	1.459e-06	3.365e-05

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.304e-04	6.637e-04	0.347	0.729
log_lag_pa_synthé	-4.072e-05	1.138e-04	-0.358	0.721
log_lag_preca_fem	2.151e-05	5.749e-05	0.374	0.709
log_lag_tpspartiel	4.184e-05	6.226e-05	0.672	0.503
log_lag_bourse	-1.579e-06	1.225e-05	-0.129	0.898
log_lag_deficit	-3.040e-06	6.551e-06	-0.464	0.644
log_lag_opi_inflation	-4.953e-05	7.497e-05	-0.661	0.511
log_lag_opi_amelio_niv_vie	-1.122e-05	1.080e-05	-1.039	0.302
log_lag_opi_env	7.365e-06	9.057e-06	0.813	0.418
log_lag_nuptialite	-3.195e-05	3.540e-05	-0.903	0.369
log_lag_spepro_cadrefemmes	4.080e-05	4.532e-05	0.900	0.371
log_lag_ivg_100	3.135e-05	5.876e-05	0.533	0.595
log_lag_étude_sup	-3.389e-05	4.934e-05	-0.687	0.494

Residual standard error: 6.32e-06 on 87 degrees of freedom

Multiple R-squared: 0.1681, Adjusted R-squared: 0.05335

F-statistic: 1.465 on 12 and 87 DF, p-value: 0.1532

```
> # Test de Breusch-Pagan
> test_bp <- bptest(model_final)
> cat("Statistique du test : ", round(test_bp$statistic, 3), "\n")
Statistique du test : 16.809
> cat("P-value : ", round(test_bp$p.value, 5), "\n")
P-value : 0.15691
> if (test_bp$p.value > 0.05) {
+   cat("H₀ est vérifiée : pas d'évidence d'hétéroscléasticité (Breusch-Pagan).\n")
+ } else {
+   cat("H₀ est rejetée : présence d'hétéroscléasticité (Breusch-Pagan).\n")
+ }
H₀ est vérifiée : pas d'évidence d'hétéroscléasticité (Breusch-Pagan).
- I
```

Test de White

```
Résumé de la régression auxiliaire (White) :
> print(summary_white)

Call:
lm(formula = residus^2 ~ poly(valeurs_ajustees, 2), data = data.frame(residus,
fitted(model_final)))

Residuals:
    Min      1Q  Median      3Q     Max 
-7.501e-06 -3.227e-06 -2.205e-06  1.614e-06  3.632e-05 

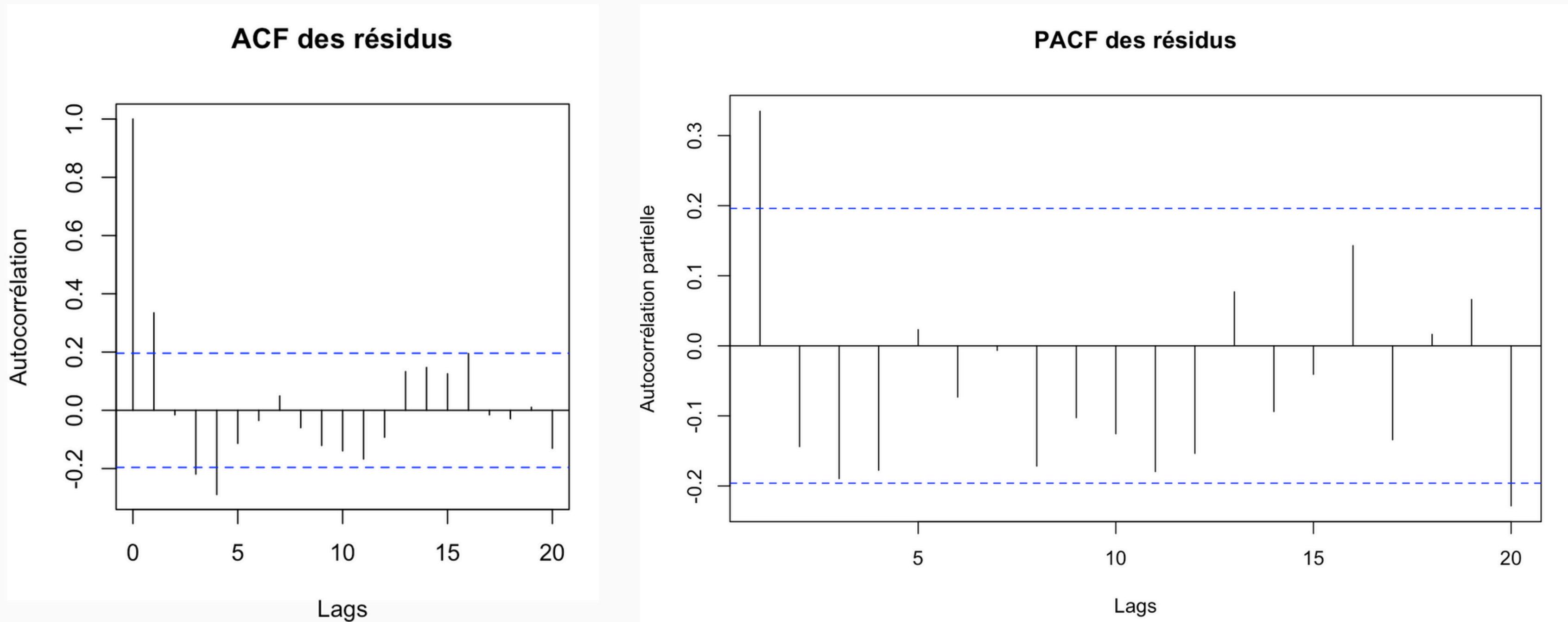
Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      4.070e-06  6.405e-07  6.354 6.81e-09 ***  
poly(valeurs_ajustees, 2)1       1.384e-05  6.405e-06  2.160  0.0332 *   
poly(valeurs_ajustees, 2)2       2.439e-06  6.405e-06  0.381  0.7042  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 6.405e-06 on 97 degrees of freedom
Multiple R-squared:  0.04725,   Adjusted R-squared:  0.0276  
F-statistic: 2.405 on 2 and 97 DF,  p-value: 0.09561

> # Statistique de White
> white_stat <- summary_white$r.squared * nrow(data_work4_log_out) # Statistique de White
> white_pval <- pchisq(white_stat, df = 2, lower.tail = FALSE) # Degré de liberté et p-value
> cat("Statistique du test :", round(white_stat, 3), "\n")
Statistique du test : 4.725
> cat("P-value :", round(white_pval, 5), "\n")
P-value : 0.09419
> if (white_pval > 0.05) {
+   cat("H0 est vérifiée : pas d'évidence d'hétéroscédasticité (White).\n")
+ } else {
+   cat("H0 est rejetée : présence d'hétéroscédasticité (White).\n")
+ }
H0 est vérifiée : pas d'évidence d'hétéroscédasticité (White).
```

Hypothèse 4 : Autocorrélation

Forme d'autocorrélation



Hypothèse 4 : Autocorrélation

Les Tests

Breusch-Godfrey test for serial correlation of order up to 40

```
data: model_final  
LM test = 60.111, df = 40, p-value = 0.02138
```

--- Détails du test de Breusch-Godfrey ---

Statistique du test : 60.1113

Degrés de liberté : 40

P-value : 0.0214

Conclusion : Autocorrélation significative détectée jusqu'à l'ordre 40 (on rejette H0).

--- Test Durbin-Watson ---

```
> dw_test <- tryCatch(dwTest(model_final), error = function(e) NULL)
```

```
> print(dw_test)
```

lag	Autocorrelation	D-W Statistic	p-value
-----	-----------------	---------------	---------

1	0.3347878	1.327747	0
---	-----------	----------	---

Alternative hypothesis: rho != 0

Durbin-Watson Statistic : 1.3277

p-value : 0

Conclusion : Les résidus suggèrent une autocorrélation positive.

Hypothèse 4 : Autocorrélation

Détection de stationnarité et de saisonnalité

```
--- Vérification de la saisonnalité ---
> seasonality_test <- kruskal.test(ts_residuals
~ cycle(ts_residuals))
> cat("Statistique du test Kruskal-Wallis :", r
ound(seasonality_test$statistic, 4), "\n")
Statistique du test Kruskal-Wallis : 0.9715
> cat("P-value :", round(seasonality_test$p.val
ue, 4), "\n")
P-value : 0.8082
> if (seasonality_test$p.value < 0.05) {
+   cat("Conclusion : Les résidus montrent une
saisonnalité significative.\n")
+ } else {
+   cat("Conclusion : Pas de saisonnalité signifi
cative détectée.\n")
+ }
Conclusion : Pas de saisonnalité significative
détectée.
```

```
--- Vérification de la stationnarité (ADF Test) ---
> adf_test <- adf.test(ts_residuals, alternative = "stationary")
Warning message:
In adf.test(ts_residuals, alternative = "stationary") :
  p-value smaller than printed p-value
> cat("Statistique ADF :", round(adf_test$statistic, 4), "\n")
Statistique ADF : -5.0416
> cat("P-value :", round(adf_test$p.value, 4), "\n")
P-value : 0.01
> if (adf_test$p.value < 0.05) {
+   cat("Conclusion : Les résidus sont stationnaires (on rejette H0).\n")
+ } else {
+   cat("Conclusion : Les résidus ne sont pas stationnaires (on ne rej
+ ")
Conclusion : Les résidus sont stationnaires (on rejette H0).
```

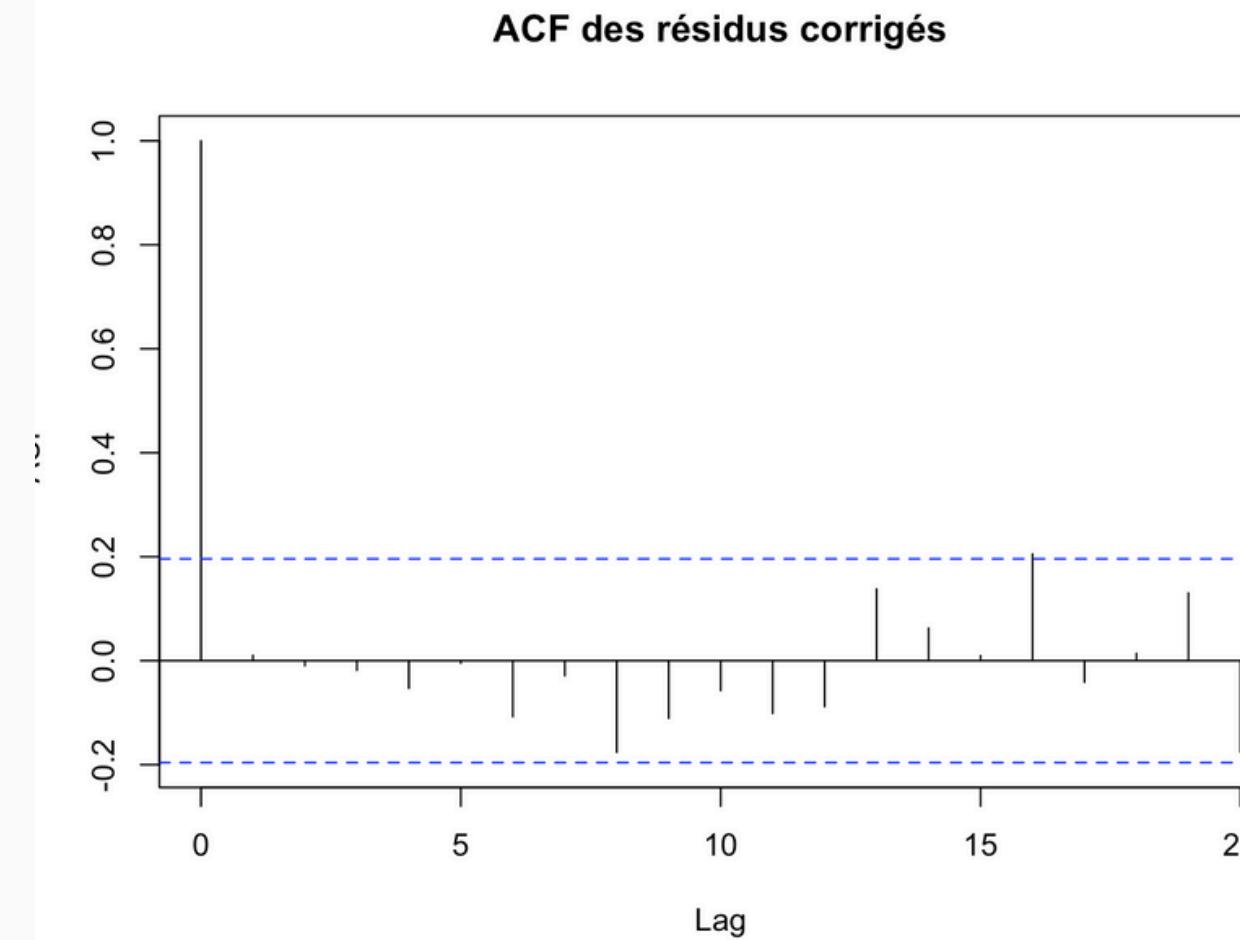
Hypothèse 4 : Autocorrélation

Solution : Modèle ARIMA (4.0.0)

```
Résumé du modèle ARIMA(4,0,0) :  
> print(summary(arima_model_400))  
  
Call:  
arima(x = residuals_final, order = c(4, 0, 0), include.mean = TRUE)  
  
Coefficients:  
      ar1      ar2      ar3      ar4  intercept  
    0.3177 -0.0803 -0.1241 -0.1764      0e+00  
  s.e.  0.0984  0.1022  0.1041  0.1017     2e-04  
  
sigma^2 estimated as 3.298e-06:  log likelihood = 489.02,  aic = -966.03  
  
Training set error measures:  
          ME        RMSE       MAE       MPE       MAPE       MASE  
Training set -4.1963e-06 0.00181617 0.001283765 -73.35124 306.9336 0.9274854  
          ACF1  
Training set 0.01042658
```

Hypothèse 4 : Autocorrélation

Solution : Modèle ARIMA (4.0.0)



```

lag Autocorrelation D-W Statistic p-value
 1      0.01042658    1.976729  0.836
Alternative hypothesis: rho != 0
> # Interprétation de la p-value
> cat("\nInterprétation :\n")

Interprétation :
> if (dw_test_corrected$p > 0.05) {
+   cat("H0 acceptée : Pas d'autocorrélation d'ordre 1 détectée.\n")
+ } else {
+   cat("H0 rejetée : Présence d'autocorrélation d'ordre 1 détectée.\n")
+ }
H0 acceptée : Pas d'autocorrélation d'ordre 1 détectée.

```

```

--- Test de Ljung-Box sur les résidus corrigés ---
> ljung_box_arima <- Box.test(arima_residuals, lag = 40, type = "Ljung-Box")
> cat("Statistique Ljung-Box :", round(ljung_box_arima$statistic, 4), "\n")
Statistique Ljung-Box : 40.7533
> cat("P-value :", round(ljung_box_arima$p.value, 4), "\n")
P-value : 0.4371
> if (ljung_box_arima$p.value > 0.05) {
+   cat("Conclusion : Les résidus du modèle ARIMA ne présentent pas d'autocorrélation significative.\n")
+ } else {
+   cat("Conclusion : Les résidus présentent encore une autocorrélation significative. Une révision du modèle ARIMA est nécessaire.\n")
+ }
Conclusion : Les résidus du modèle ARIMA ne présentent pas d'autocorrélation significative.

```

Notre Modèle Final

Nouvelle régression OLS avec résidus corrigés de l'autocorrélation

```

Residuals:
    Min      1Q   Median     3Q     Max
-0.0058845 -0.0011722  0.0001364  0.0013188  0.0065064

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                         0.216308  0.227131  0.952  0.343559
log_lag_pa_synthé                  -0.174425  0.038946 -4.479 2.27e-05 ***
log_lag_preca_fem                 -0.013708  0.019677 -0.697  0.487867
log_lag_tpspartiel                0.060950  0.021308  2.860  0.005296 **
log_lag_bourse                     0.019317  0.004193  4.607 1.39e-05 ***
log_lag_deficit                   0.013926  0.002242  6.211 1.75e-08 ***
log_lag_opi_inflation              0.066211  0.025657  2.581  0.011538 *
log_lag_opi_amelio_niv_vie       0.014455  0.003696  3.911  0.000182 ***
log_lag_opi_env                   0.007718  0.003100  2.490  0.014671 *
log_lag_nuptialite               -0.007631  0.012115 -0.630  0.530444
log_lag_spepro_cadrefemmes        0.061821  0.015511  3.986  0.000140 ***
log_lag_ivg_100                   0.032391  0.020110  1.611  0.110872
log_lag_étude_sup                -0.036704  0.016885 -2.174  0.032436 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002163 on 87 degrees of freedom
Multiple R-squared:  0.7653,    Adjusted R-squared:  0.7329
F-statistic: 23.64 on 12 and 87 DF,  p-value: < 2.2e-16

```

Les Limites



Généralisation de nos résultats



Nombre faibles d'années d'observation. Impact de l'interpolation



Nos hypothèses d'endogénéité (subjectives et basées sur nos opinions et connaissances en économie)