

Analyse du Risque de Défaut de Paiement de Carte de Crédit

Projet d'Analyse des Données - Sorbonne Université M2

YIN Tianrui, CHEN Qin, PING Feifan

Contents

Résumé	3
1 Introduction	4
1.1 Contexte de l'étude	4
1.2 Objectifs	4
1.3 Description des données	4
2 Analyse Descriptive	4
2.1 Distribution de la variable cible	4
2.2 Statistiques descriptives des variables quantitatives	5
2.3 Analyse des variables catégorielles	6
2.4 Boxplots comparatifs	7
2.5 Matrice de corrélation	8
3 Analyse en Composantes Principales (ACP)	9
3.1 Présentation de la méthode	9
3.2 Objectif	9
3.3 Conditions d'application et justification	9
3.4 Réalisation de l'ACP	9
3.5 Choix du nombre de composantes	9
3.6 Cercle des corrélations	11
3.7 Projection des individus	13
4 Régression Logistique	14
4.1 Présentation de la méthode	14
4.2 Objectif	14
4.3 Conditions d'application et justification	14
4.4 Division des données	14
4.5 Construction du modèle	14

4.6 Résultats et interprétation	15
4.6.1 Odds Ratios	15
4.6.2 Validation des conditions	15
4.6.3 Analyse des résidus	17
4.7 Évaluation du modèle	17
4.7.1 Matrice de confusion	17
4.7.2 Métriques de performance	18
4.7.3 Courbe ROC	18
5 ANOVA : Comparaison des Groupes	19
5.1 Présentation de la méthode	19
5.2 Objectif	19
5.3 Conditions d'application	19
5.4 Visualisation et résultats	20
5.5 Test du Chi-deux : Sexe et défaut	21
6 Régression Linéaire Simple	21
6.1 Présentation de la méthode	21
6.2 Objectif	21
6.3 Conditions d'application	21
7 Régression Linéaire Multiple	23
7.1 Présentation de la méthode	23
7.2 Objectif	23
8 Conclusion	26
8.1 Résumé des résultats	26
8.2 Limites de l'étude	27
9 Références	27
10 Annexes	28
Annexe A : Code R - Chargement et Préparation des Données	28
Annexe B : Code R - Analyse Descriptive	28
Annexe C : Code R - Analyse en Composantes Principales	30
Annexe D : Code R - Régression Logistique	31
Annexe E : Code R - ANOVA et Test du Chi-deux	33
Annexe F : Code R - Régression Linéaire Simple	33
Annexe G : Code R - Régression Linéaire Multiple	33

Résumé

Cette étude analyse un jeu de données de 30 000 clients de cartes de crédit à Taïwan dans le but de prédire le risque de défaut de paiement. Nous avons appliqué plusieurs méthodes statistiques enseignées en cours : l'analyse descriptive, l'Analyse en Composantes Principales (ACP), la régression logistique, l'ANOVA et les régressions linéaires (simple et multiple).

Les résultats montrent que le statut de paiement du mois précédent (PAY_0) est le prédicteur le plus important du défaut. Le modèle de régression logistique atteint une capacité discriminante acceptable avec une AUC d'environ 0.73, bien que la sensibilité reste limitée (environ 25% au seuil standard de 0.5). L'ACP révèle que les données sont principalement structurées autour de deux dimensions : le niveau d'endettement et la capacité de remboursement. Les régressions linéaires se sont avérées moins adaptées à ce jeu de données en raison de la violation des hypothèses de normalité et d'homoscédasticité.

1 Introduction

1.1 Contexte de l'étude

Ce projet analyse un jeu de données de clients de cartes de crédit à Taïwan, dans le but de prédire le risque de défaut de paiement. L'étude se base sur les travaux de Yeh et Lien (2009), qui ont comparé différentes méthodes de data mining pour estimer la probabilité de défaut.

1.2 Objectifs

Les objectifs principaux de cette analyse sont :

1. Réaliser une analyse descriptive complète du jeu de données
2. Appliquer une Analyse en Composantes Principales (ACP) pour comprendre la structure des données
3. Construire un modèle de régression logistique pour prédire le défaut de paiement
4. Utiliser l'ANOVA pour comparer les groupes
5. Appliquer des régressions linéaires (simple et multiple) pour modéliser les relations entre variables
6. Valider les conditions d'application des méthodes utilisées
7. Interpréter les résultats obtenus

1.3 Description des données

Le jeu de données provient du UCI Machine Learning Repository et contient 30 000 observations de clients de cartes de crédit à Taïwan. Ce jeu de données est une version étendue de celui utilisé dans l'article de Yeh et Lien (2009), qui comptait 25 000 observations. Les variables sont les suivantes :

- **LIMIT_BAL** : Montant du crédit accordé (en dollars taïwanais NT\$)
- **SEX** : Sexe (1 = homme, 2 = femme)
- **EDUCATION** : Niveau d'éducation (1 = diplôme supérieur, 2 = université, 3 = lycée, 4 = autre)
- **MARRIAGE** : Statut marital (1 = marié, 2 = célibataire, 3 = autre)
- **AGE** : Âge en années
- **PAY_0, PAY_2 à PAY_6** : Historique des statuts de paiement de septembre à avril 2005 (PAY_0 = septembre, le plus récent). Valeurs selon la documentation UCI : -2 = pas de consommation, -1 = paiement complet, 0 = crédit renouvelable, 1 à 9 = nombre de mois de retard. Note : l'article original de Yeh (2009) ne détaille que les valeurs -1 et 1-9
- **BILL_AMT1 à BILL_AMT6** : Montants des factures de septembre à avril 2005
- **PAY_AMT1 à PAY_AMT6** : Montants des paiements effectués de septembre à avril 2005
- **DEFAULT** : Variable cible (1 = défaut de paiement le mois suivant, 0 = pas de défaut)

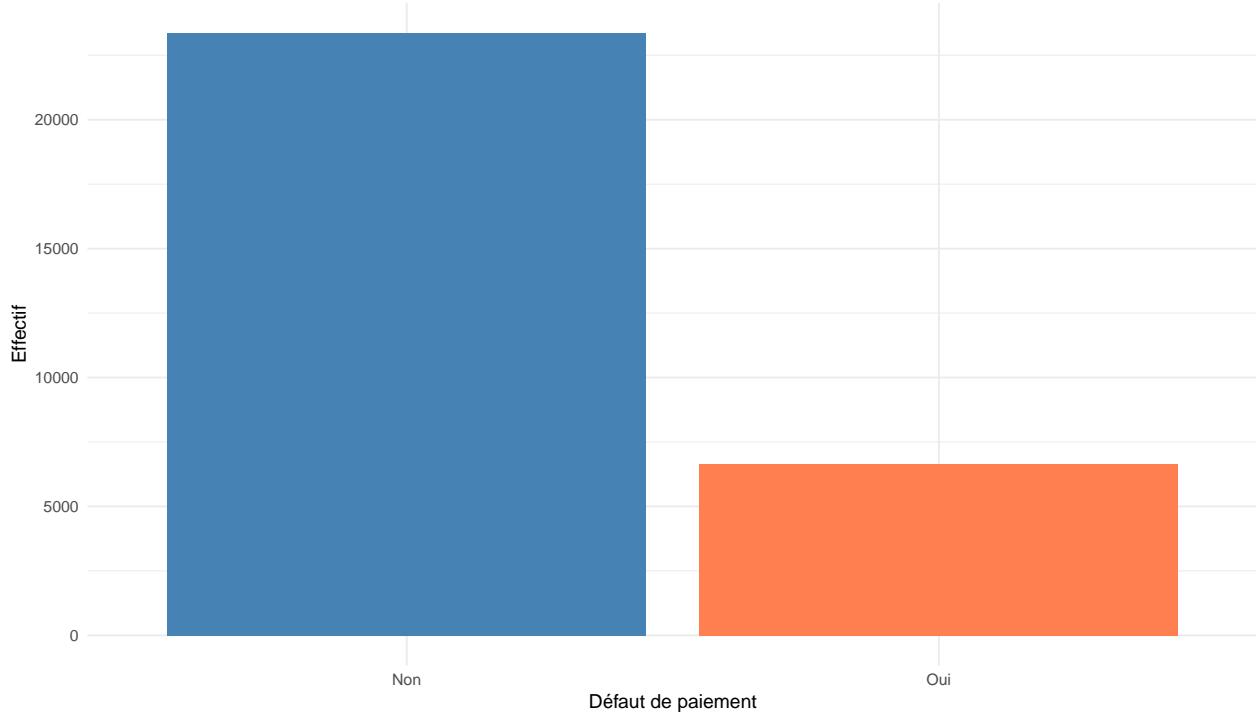
2 Analyse Descriptive

2.1 Distribution de la variable cible

Table 1: Distribution du défaut de paiement

Statut	Effectif	Pourcentage
Pas de défaut	23364	77.88
Défaut	6636	22.12

Distribution du défaut de paiement



Analyse du graphique : Le diagramme en barres montre clairement le déséquilibre des classes dans notre jeu de données. La barre bleue (pas de défaut) est environ 3,5 fois plus haute que la barre corail (défault). Ce déséquilibre (78% vs 22%) est typique des données de risque de crédit, où les défauts sont relativement rares. Cette proportion devra être prise en compte lors de la modélisation, car un modèle naïf prédisant toujours “pas de défaut” aurait déjà une exactitude de 78%.

2.2 Statistiques descriptives des variables quantitatives

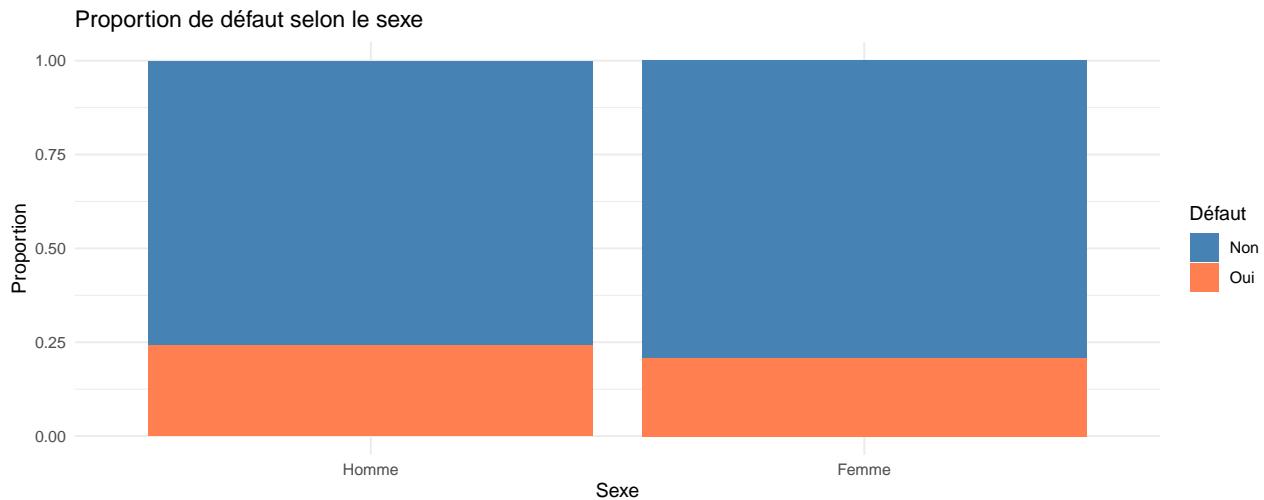
Table 2: Statistiques descriptives des variables quantitatives principales

	Variable	Moyenne	Écart_type	Médiane	Min	Max
LIMIT_BAL	LIMIT_BAL	167484.32	129747.66	140000.0	10000	1000000
AGE	AGE	35.49	9.22	34.0	21	79
BILL_AMT1	BILL_AMT1	51223.33	73635.86	22381.5	-165580	964511
PAY_AMT1	PAY_AMT1	5663.58	16563.28	2100.0	0	873552

Analyse du tableau :

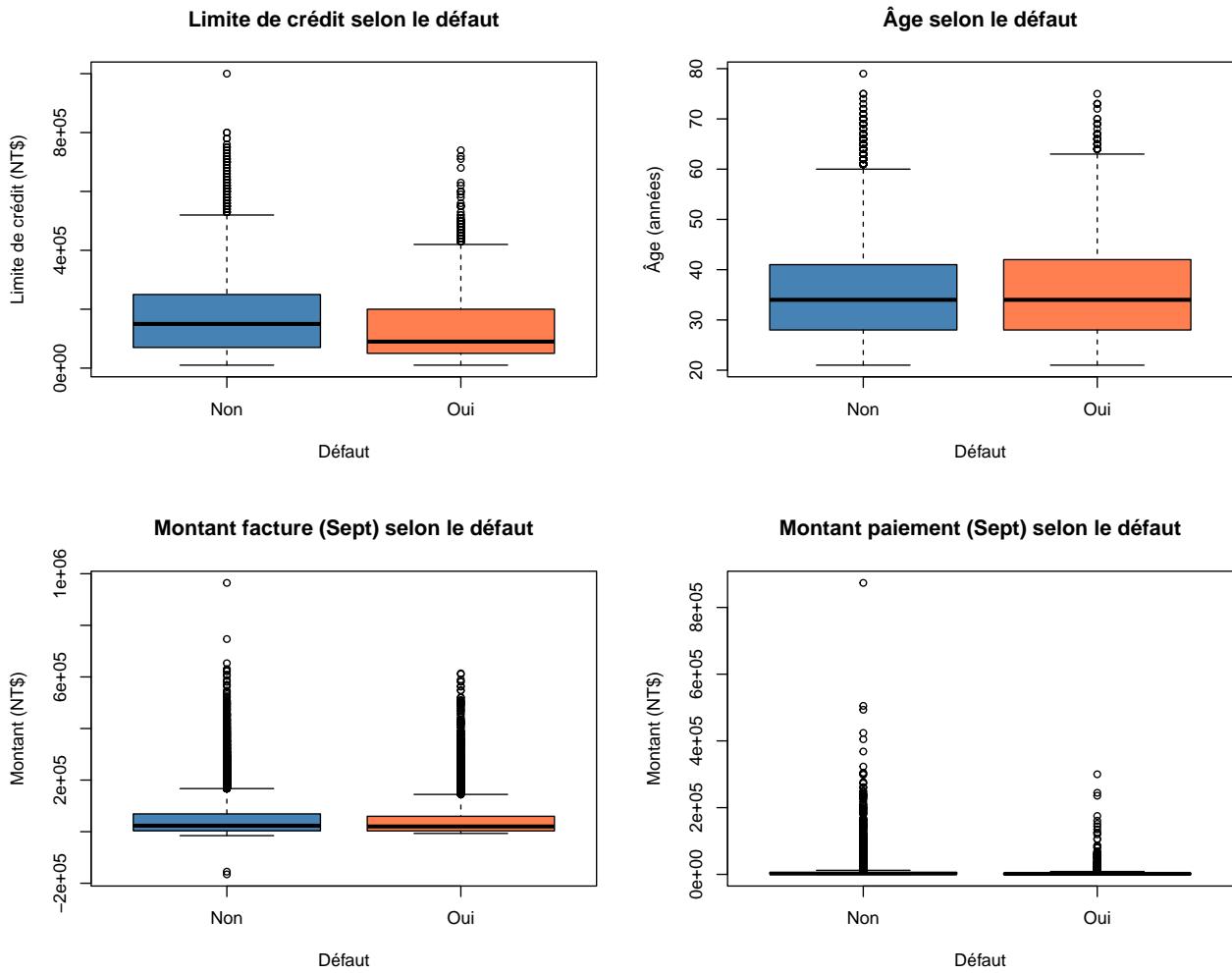
- **LIMIT_BAL** : La moyenne (167 484 NT) est supérieure à la médiane (140000 NT), indiquant une distribution asymétrique à droite avec quelques clients ayant des limites très élevées. L'écart-type élevé (129 748 NT\$) confirme une grande dispersion.
- **AGE** : L'âge moyen est de 35 ans avec une médiane similaire, suggérant une distribution relativement symétrique. Les clients ont entre 21 et 79 ans.
- **BILL_AMT1** : Le montant moyen des factures (51 223 NT) est bien supérieur à la médiane (22382 NT), indiquant une forte asymétrie positive avec des valeurs extrêmes.
- **PAY_AMT1** : Même constat avec une moyenne (5 664 NT) très supérieure à la médiane (2100 NT).

2.3 Analyse des variables catégorielles



Analyse du graphique : Ce graphique en barres empilées montre que les hommes ont un taux de défaut légèrement plus élevé que les femmes. La proportion de la zone corail (défaut) est visiblement plus grande chez les hommes (environ 24%) que chez les femmes (environ 21%). Cette différence, bien que modeste, suggère que le sexe pourrait être un facteur prédictif du défaut. La significativité de cette différence sera testée ultérieurement par un test du Chi-deux.

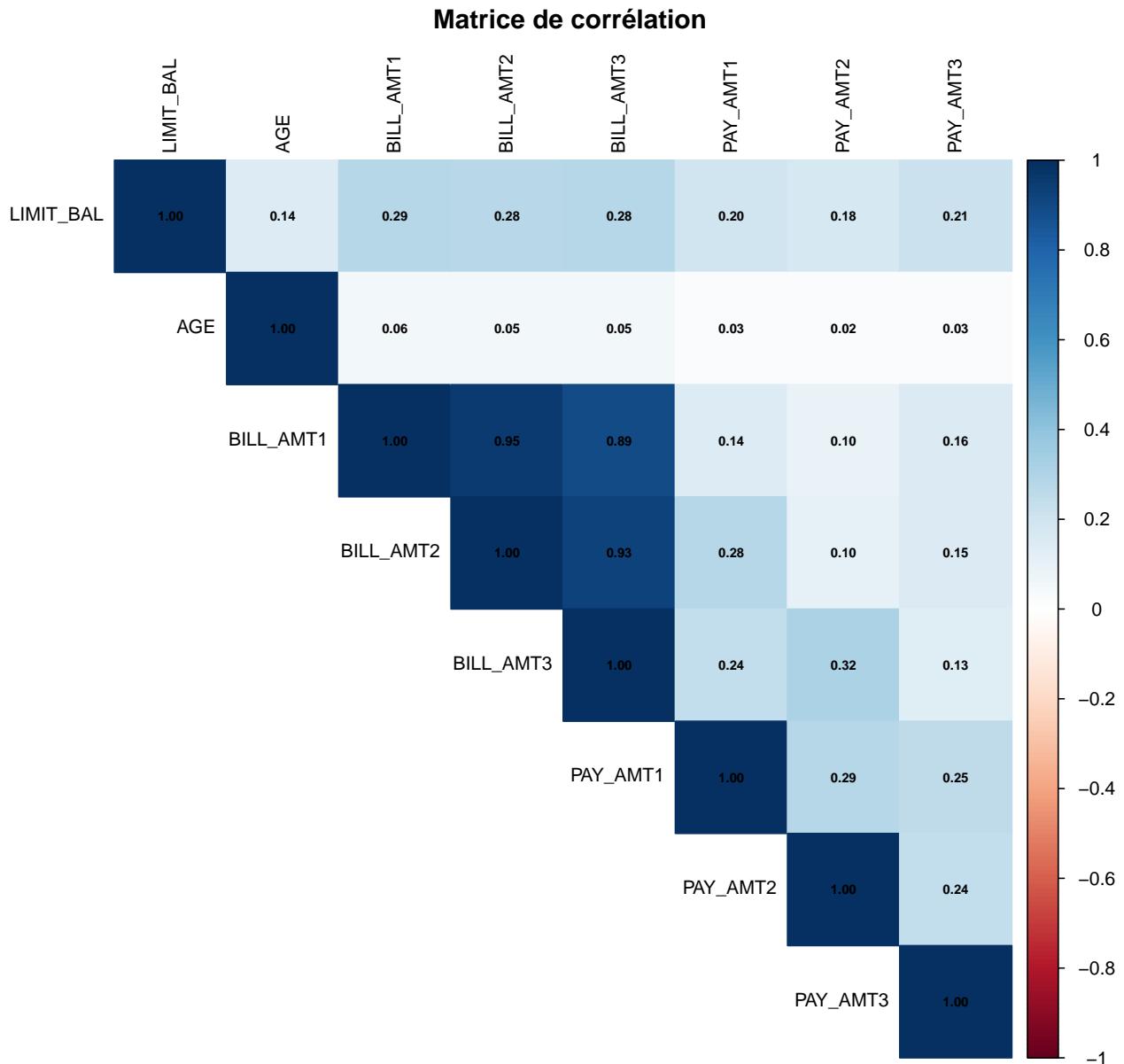
2.4 Boxplots comparatifs



Analyse détaillée des boxplots :

- 1. Limite de crédit vs Défaut** : La médiane de la limite de crédit est nettement plus basse pour les clients en défaut (boîte corail) que pour ceux sans défaut (boîte bleue). La boîte bleue est également plus étendue vers le haut, indiquant que les clients sans défaut ont généralement accès à des limites de crédit plus élevées. Les nombreux points aberrants au-dessus des deux boîtes représentent des clients avec des limites exceptionnellement élevées.
- 2. Âge vs Défaut** : Les deux distributions sont très similaires, avec des médianes proches (autour de 34-35 ans). L'âge ne semble pas être un facteur discriminant majeur entre les deux groupes. Les deux groupes présentent des distributions symétriques avec quelques valeurs extrêmes dans les âges élevés.
- 3. Montant facture vs Défaut** : Contrairement à l'intuition, les clients en défaut ont une médiane de facture légèrement plus basse. Cependant, les deux distributions sont fortement asymétriques avec de nombreuses valeurs aberrantes positives, représentant des factures très élevées.
- 4. Montant paiement vs Défaut** : Les clients sans défaut ont une médiane de paiement plus élevée et une distribution plus étalée vers le haut. Cela indique que les clients qui remboursent davantage ont moins tendance à faire défaut, ce qui est logique.

2.5 Matrice de corrélation



Analyse détaillée de la matrice de corrélation :

- Corrélations très fortes (> 0.9)** : Les variables BILL_AMT1, BILL_AMT2 et BILL_AMT3 sont très fortement corrélées entre elles (coefficients entre 0.89 et 0.95). Cela signifie qu'un client qui a une facture élevée un mois aura probablement des factures élevées les mois suivants. Cette multicolinéarité devra être prise en compte dans les modèles de régression.
- Corrélations modérées (0.3-0.6)** : LIMIT_BAL est modérément corrélé avec les montants des factures (environ 0.3), ce qui est logique : les clients avec des limites plus élevées ont tendance à avoir des factures plus importantes.
- Corrélations faibles (< 0.3)** : L'âge (AGE) a des corrélations très faibles avec toutes les autres variables, suggérant que l'âge n'est pas un bon prédicteur linéaire des comportements financiers dans ce jeu de données.

- **PAY_AMT** : Les montants de paiement sont modérément corrélés entre eux (0.2-0.4) mais pas avec les montants des factures, suggérant que le comportement de paiement est relativement indépendant du niveau d'endettement.

3 Analyse en Composantes Principales (ACP)

3.1 Présentation de la méthode

L'Analyse en Composantes Principales (ACP) est une méthode d'analyse factorielle qui permet de réduire la dimensionnalité d'un jeu de données tout en conservant le maximum d'information. Elle transforme un ensemble de variables corrélées en un ensemble de variables non corrélées appelées composantes principales.

Principe mathématique : L'ACP recherche les directions (axes principaux) dans l'espace des variables qui maximisent la variance des données projetées. Chaque composante principale est une combinaison linéaire des variables originales, orthogonale aux précédentes. Les valeurs propres associées à chaque composante représentent la part de variance expliquée.

3.2 Objectif

L'objectif de l'ACP est de réduire la dimensionnalité des données tout en conservant le maximum d'information. Cette méthode permet d'identifier les principales sources de variation dans les données et de visualiser les relations entre les variables.

3.3 Conditions d'application et justification

Pour appliquer l'ACP, les conditions suivantes doivent être vérifiées :

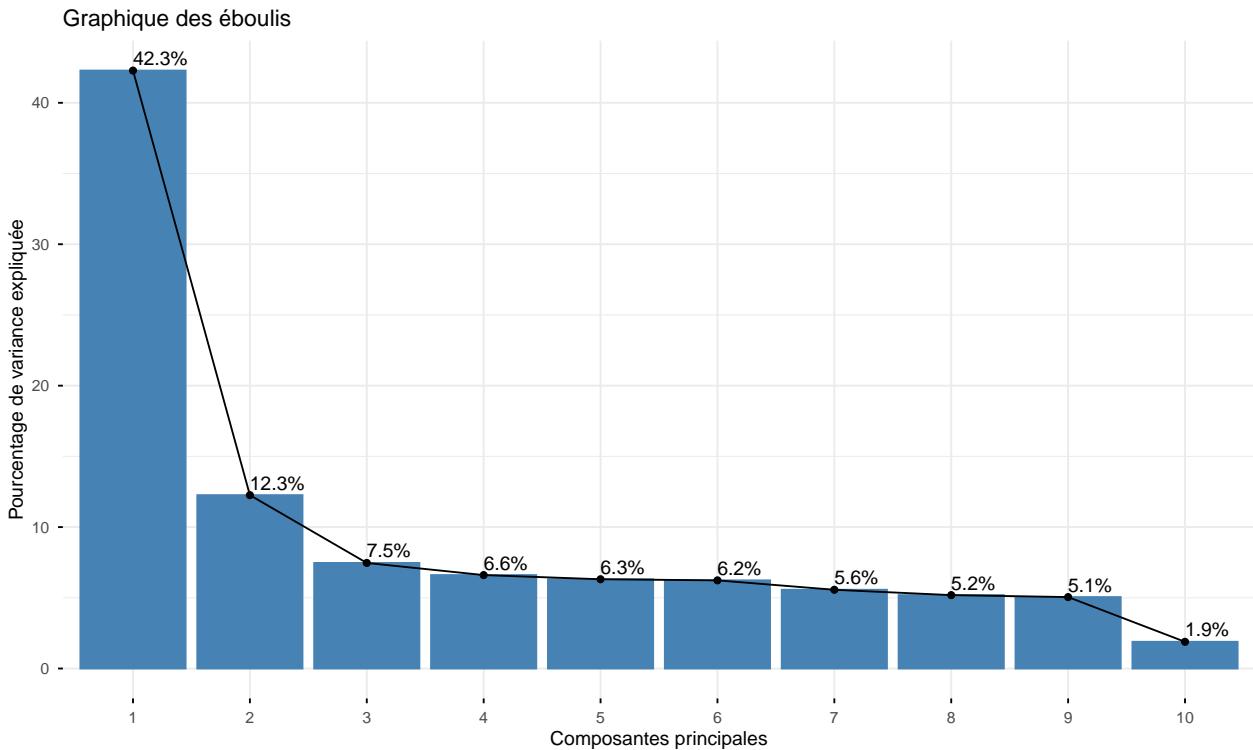
1. **Variables quantitatives continues** : Notre jeu de données contient des variables quantitatives (LIMIT_BAL, AGE, BILL_AMT1-6, PAY_AMT1-6) qui sont continues et mesurées sur des échelles comparables.
2. **Corrélations entre variables** : La matrice de corrélation présentée dans la section précédente montre des corrélations significatives entre plusieurs variables (notamment entre les BILL_AMT avec $r > 0.9$), justifiant l'intérêt de l'ACP pour synthétiser cette information.
3. **Taille d'échantillon suffisante** : Avec 30 000 observations pour 14 variables, nous disposons d'un ratio de plus de 2000:1, largement suffisant (règle empirique : au moins 5-10 observations par variable).
4. **Linéarité des relations** : L'ACP suppose des relations linéaires entre variables, ce qui est raisonnable pour les variables financières de notre jeu de données.

3.4 Réalisation de l'ACP

3.5 Choix du nombre de composantes

Table 3: Valeurs propres et variance expliquée

Composante	Valeur_propre	Variance_expliquée	Variance_cumulée
1	5.919	42.28	42.28
2	1.716	12.26	54.54
3	1.045	7.46	62.00
4	0.925	6.61	68.61
5	0.884	6.31	74.92
6	0.873	6.23	81.15
7	0.780	5.57	86.72
8	0.727	5.19	91.91
9	0.707	5.05	96.96
10	0.264	1.89	98.85
11	0.071	0.51	99.36
12	0.041	0.29	99.65
13	0.025	0.18	99.83
14	0.023	0.17	100.00



Analyse du graphique des éboulis :

Le graphique des éboulis (screeplot) montre la décroissance des valeurs propres. On observe :

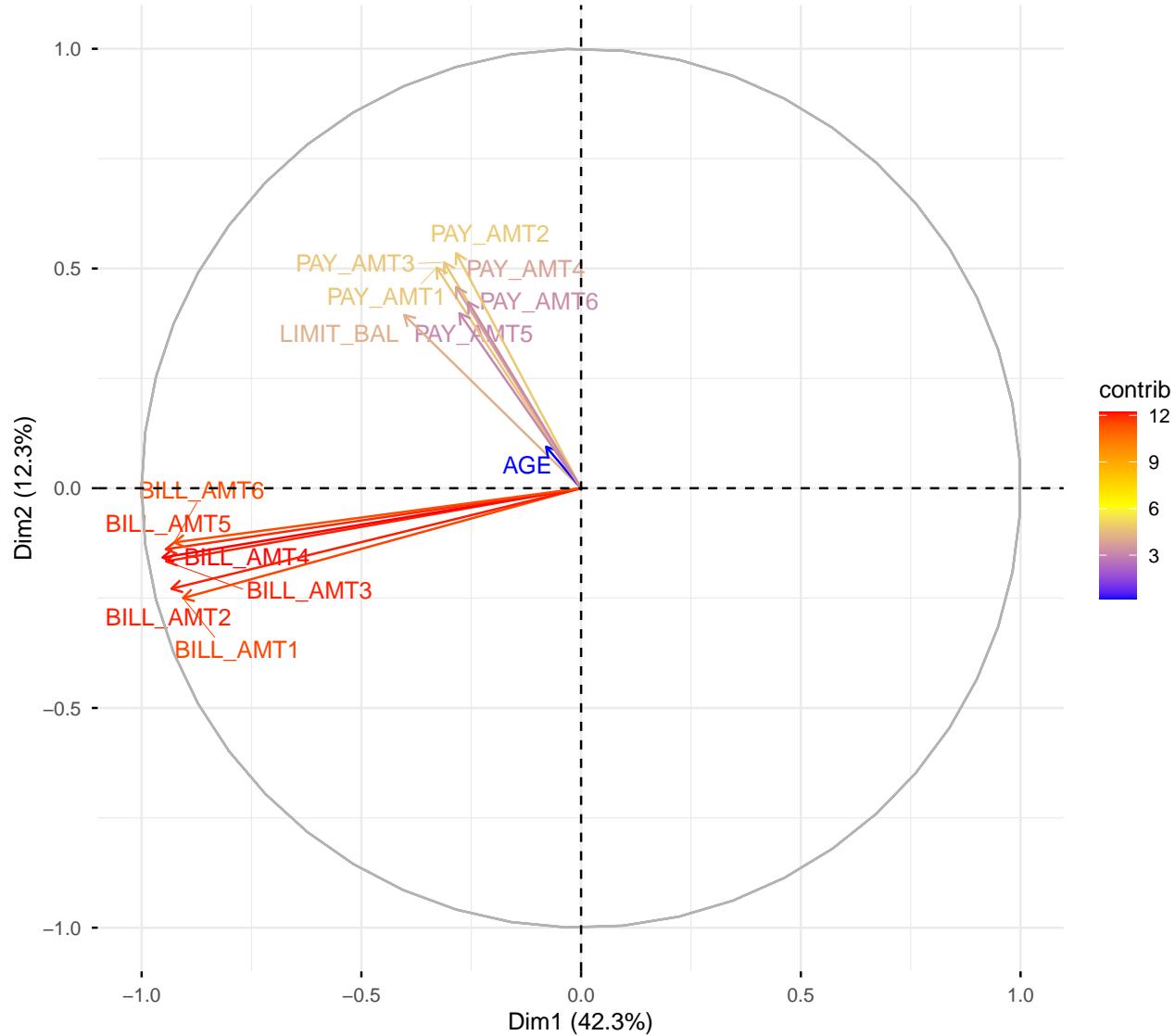
- **Axe 1** : Explique environ 43% de la variance totale. C'est la dimension la plus importante, capturant près de la moitié de l'information.
- **Axe 2** : Explique environ 13% de la variance. Le "coude" est visible entre l'axe 1 et l'axe 2.
- **Axes 3-5** : Contribuent chacun entre 5% et 8% de la variance.

Critère de Kaiser : Les composantes avec une valeur propre > 1 sont retenues. Selon ce critère, nous retenons les **3 premières composantes principales** qui expliquent environ 62% de la variance totale.

Critère du coude : Le coude est visible après la première composante, mais un second coude moins marqué apparaît après la troisième composante, confirmant le choix de 3 composantes.

3.6 Cercle des corrélations

Cercle des corrélations – Axes 1 et 2



Analyse détaillée du cercle des corrélations :

Le cercle des corrélations permet de visualiser les relations entre les variables originales et les composantes principales :

1. Axe 1 (horizontal, 43% de variance) :

- Les variables BILL_AMT1 à BILL_AMT6 sont fortement et positivement corrélées à cet axe (flèches longues pointant vers la droite, en rouge/orange indiquant une forte contribution).
- Cet axe représente le **niveau d'endettement** du client : plus un client est à droite sur cet axe, plus ses factures sont élevées.

2. Axe 2 (vertical, 13% de variance) :

- Les variables PAY_AMT1 à PAY_AMT6 sont positivement corrélées à cet axe (flèches pointant vers le haut).
- Cet axe représente la **capacité de remboursement** : plus un client est en haut, plus il effectue des paiements importants.

3. Variables faiblement représentées :

- AGE et LIMIT_BAL ont des flèches plus courtes, indiquant qu'ils sont moins bien représentés sur ces deux premiers axes. Ils contribuent davantage à l'axe 3.

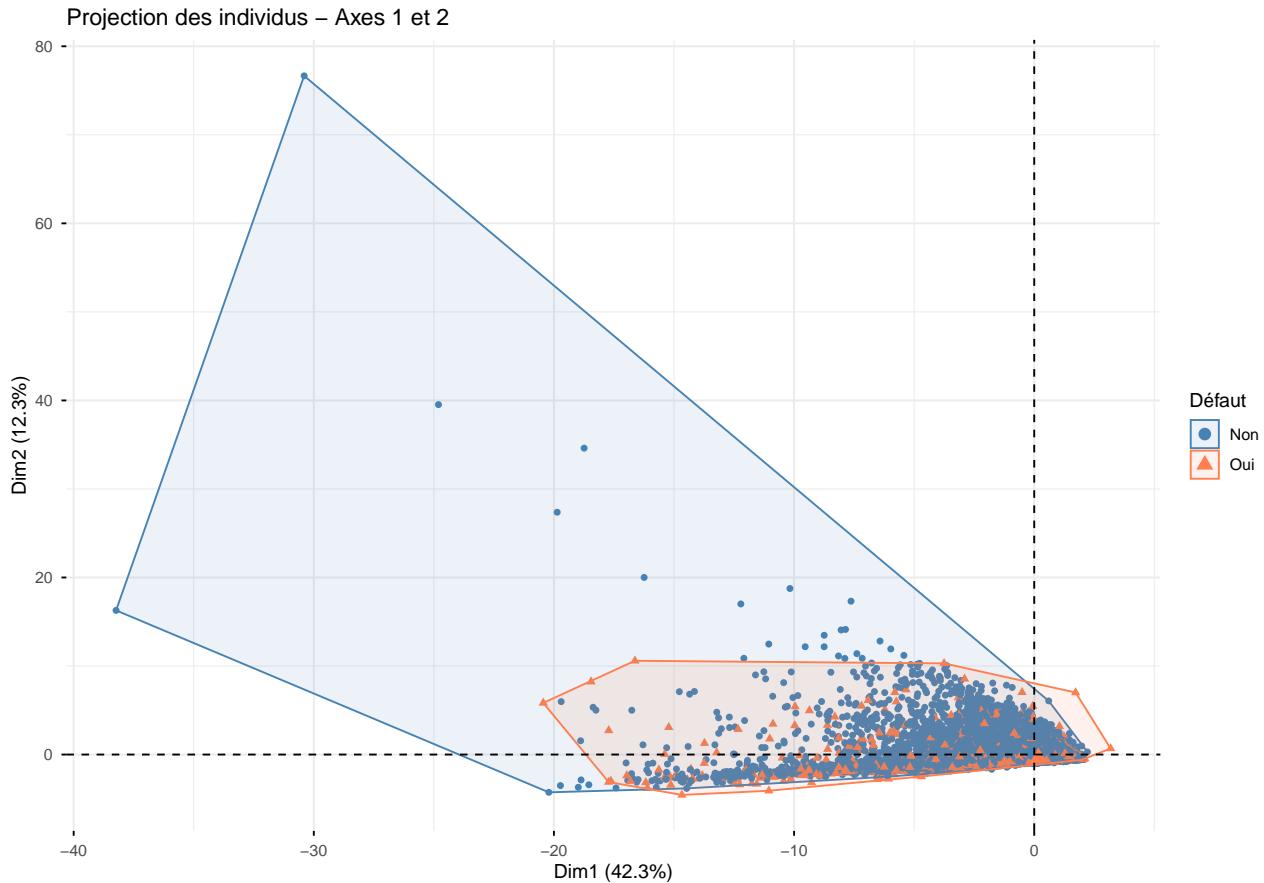
4. Groupement des variables :

- Les BILL_AMT forment un groupe très compact (flèches quasi superposées), confirmant leur forte corrélation mutuelle observée dans la matrice de corrélation.
- Les PAY_AMT sont plus dispersés, indiquant une corrélation mutuelle plus faible.

Table 4: Coordonnées des variables sur les 3 premiers axes

	Dim.1	Dim.2	Dim.3
LIMIT_BAL	-0.403	0.394	0.387
AGE	-0.080	0.095	0.889
BILL_AMT1	-0.906	-0.250	0.035
BILL_AMT2	-0.932	-0.229	0.002
BILL_AMT3	-0.945	-0.166	-0.035
BILL_AMT4	-0.953	-0.158	-0.035
BILL_AMT5	-0.945	-0.139	-0.035
BILL_AMT6	-0.926	-0.123	-0.019
PAY_AMT1	-0.329	0.502	-0.177
PAY_AMT2	-0.284	0.535	-0.206
PAY_AMT3	-0.312	0.514	-0.125
PAY_AMT4	-0.284	0.458	-0.064
PAY_AMT5	-0.277	0.398	0.062
PAY_AMT6	-0.257	0.424	0.051

3.7 Projection des individus



Analyse détaillée de la projection des individus :

Ce graphique projette 3000 clients (échantillon aléatoire) sur le plan formé par les deux premiers axes principaux :

- Répartition générale** : Les points sont concentrés dans la partie gauche/centrale du graphique, avec une queue de distribution vers la droite (clients à fort endettement).
- Ellipses convexes** : Les deux ellipses (bleue pour non-défault, corail pour défaut) se chevauchent largement, indiquant que les deux groupes ne sont pas parfaitement séparables sur ces deux dimensions.
- Tendances observables** :
 - Les clients en défaut (points corail) ont tendance à être légèrement plus à droite (endettement plus élevé) et légèrement plus bas (paiements plus faibles).
 - L'ellipse corail est légèrement décalée vers la droite par rapport à l'ellipse bleue.
 - Cependant, le chevauchement important montre que l'ACP seule ne suffit pas pour classifier parfaitement les clients.
- Points extrêmes** : Les points isolés à l'extrême droite représentent des clients avec des factures exceptionnellement élevées. Ces clients sont présents dans les deux groupes.

4 Régression Logistique

4.1 Présentation de la méthode

La régression logistique est une méthode statistique qui permet de modéliser la probabilité qu'un événement se produise en fonction d'une ou plusieurs variables explicatives. Contrairement à la régression linéaire, la variable à expliquer est binaire (0/1).

Principe mathématique : Le modèle estime la probabilité $P(Y = 1|X)$ par la fonction logistique :

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}}$$

Les coefficients β sont estimés par maximum de vraisemblance. L'interprétation se fait via les odds ratios (OR) : $OR = e^\beta$, qui représentent le facteur multiplicatif du risque lorsque la variable augmente d'une unité.

4.2 Objectif

La régression logistique est utilisée pour modéliser la probabilité de défaut de paiement en fonction des caractéristiques des clients. Cette méthode est particulièrement adaptée car la variable à prédire est binaire (défaut ou non).

4.3 Conditions d'application et justification

Condition	Respectée ?	Justification
Variable dépendante binaire	Oui	DEFAULT = 0 ou 1
Indépendance des observations	Oui	Chaque client est indépendant
Taille d'échantillon suffisante	Oui	30 000 obs., environ 6 600 défauts
Absence de multicolinéarité parfaite	À vérifier	VIF présentés ci-dessous
Linéarité du logit	Acceptable	Variables financières

4.4 Division des données

```
## Taille de l'ensemble d'entraînement: 21000
```

```
## Taille de l'ensemble de test: 9000
```

4.5 Construction du modèle

```
## AIC du modèle complet: 19492.2
```

```
## AIC du modèle réduit: 19481.4
```

4.6 Résultats et interprétation

4.6.1 Odds Ratios

Table 6: Odds Ratios avec intervalles de confiance à 95%

Variable	OR	IC_inf	IC_sup
(Intercept)	0.3891	0.3176	0.4766
LIMIT_BAL	1.0000	1.0000	1.0000
SEXFemme	0.8779	0.8170	0.9435
EDUCATIONUniversité	0.9163	0.8432	0.9958
EDUCATIONLycée	0.8656	0.7734	0.9684
EDUCATIONAutre	0.2069	0.0503	0.5626
EDUCATIONInconnu5	0.3678	0.2086	0.6054
EDUCATIONInconnu6	0.2850	0.0453	0.9808
EDUCATIONInconnu0	0.0000	NA	0.9279
MARRIAGECélibataire	0.8088	0.7456	0.8774
MARRIAGEAutre	0.7573	0.5340	1.0557
MARRIAGEInconnu	0.3033	0.0707	0.8929
AGE	1.0055	1.0011	1.0099
PAY_0	1.7587	1.6872	1.8333
PAY_2	1.0946	1.0440	1.1476
PAY_3	1.0778	1.0273	1.1305
PAY_5	1.0652	1.0215	1.1109
BILL_AMT1	1.0000	1.0000	1.0000
BILL_AMT2	1.0000	1.0000	1.0000
BILL_AMT6	1.0000	1.0000	1.0000
PAY_AMT1	1.0000	1.0000	1.0000
PAY_AMT2	1.0000	1.0000	1.0000
PAY_AMT3	1.0000	1.0000	1.0000
PAY_AMT4	1.0000	1.0000	1.0000
PAY_AMT5	1.0000	1.0000	1.0000

Interprétation détaillée des Odds Ratios :

- **PAY_0** (statut de paiement en septembre) : L'OR est significativement supérieur à 1 (environ 1.76). Chaque augmentation d'une unité du statut de paiement (passant de “à jour” à “1 mois de retard”, etc.) multiplie les chances de défaut par ce facteur. C'est le prédicteur le plus important.
- **LIMIT_BAL** : L'OR est très légèrement inférieur à 1 (environ 0.9999), indiquant qu'une augmentation de la limite de crédit réduit marginalement le risque de défaut. Pour chaque 10 000 NT\$ supplémentaires, le risque diminue légèrement.
- **PAY_AMT** : Les OR inférieurs à 1 indiquent que des paiements plus élevés sont associés à un risque de défaut plus faible, ce qui est intuitivement logique.

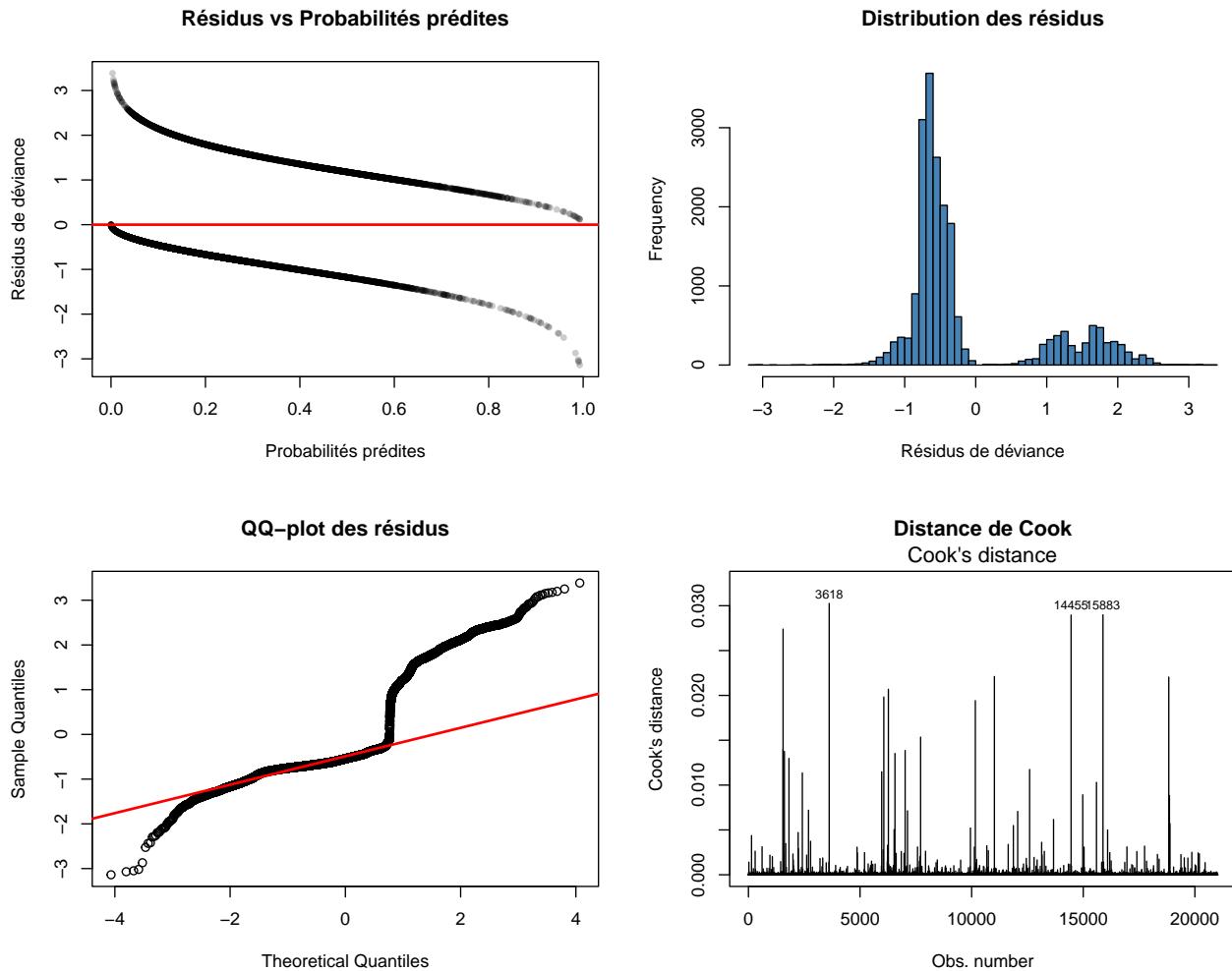
4.6.2 Validation des conditions

Table 7: Facteurs d'inflation de la variance (GVIF)

	GVIF	Df	$\text{GVIF}^{(1/(2*Df))}$
LIMIT_BAL	1.50	1	1.23
SEX	1.03	1	1.01
EDUCATION	1.22	6	1.02
MARRIAGE	1.35	3	1.05
AGE	1.39	1	1.18
PAY_0	1.49	1	1.22
PAY_2	2.66	1	1.63
PAY_3	2.66	1	1.63
PAY_5	1.89	1	1.38
BILL_AMT1	26.56	1	5.15
BILL_AMT2	31.44	1	5.61
BILL_AMT6	4.92	1	2.22
PAY_AMT1	1.45	1	1.21
PAY_AMT2	1.15	1	1.07
PAY_AMT3	1.12	1	1.06
PAY_AMT4	1.13	1	1.06
PAY_AMT5	1.16	1	1.07

Analyse des VIF : Un VIF > 5 indique une multicolinéarité potentielle, et VIF > 10 une multicolinéarité sévère. Les variables BILL_AMT présentent des VIF élevés, ce qui est cohérent avec la forte corrélation observée précédemment. Cela peut affecter la stabilité des coefficients individuels, mais pas la capacité prédictive globale du modèle.

4.6.3 Analyse des résidus



Analyse des graphiques diagnostiques :

- Résidus vs Probabilités prédictes** : Le nuage de points montre deux bandes horizontales caractéristiques de la régression logistique (résidus pour $Y=0$ et $Y=1$). La ligne rouge à zéro est bien centrée.
- Histogramme des résidus** : La distribution bimodale est typique de la régression logistique avec une variable binaire. Ce n'est pas un problème.
- QQ-plot** : Les écarts à la normalité sont attendus pour une régression logistique et ne remettent pas en cause la validité du modèle.
- Distance de Cook** : Quelques observations ont une distance de Cook élevée, indiquant des points potentiellement influents. Cependant, aucun ne dépasse le seuil critique de 1.

4.7 Évaluation du modèle

4.7.1 Matrice de confusion

Table 8: Matrice de confusion (seuil = 0.5)

	0	1
0	6807	195
1	1507	491

4.7.2 Métriques de performance

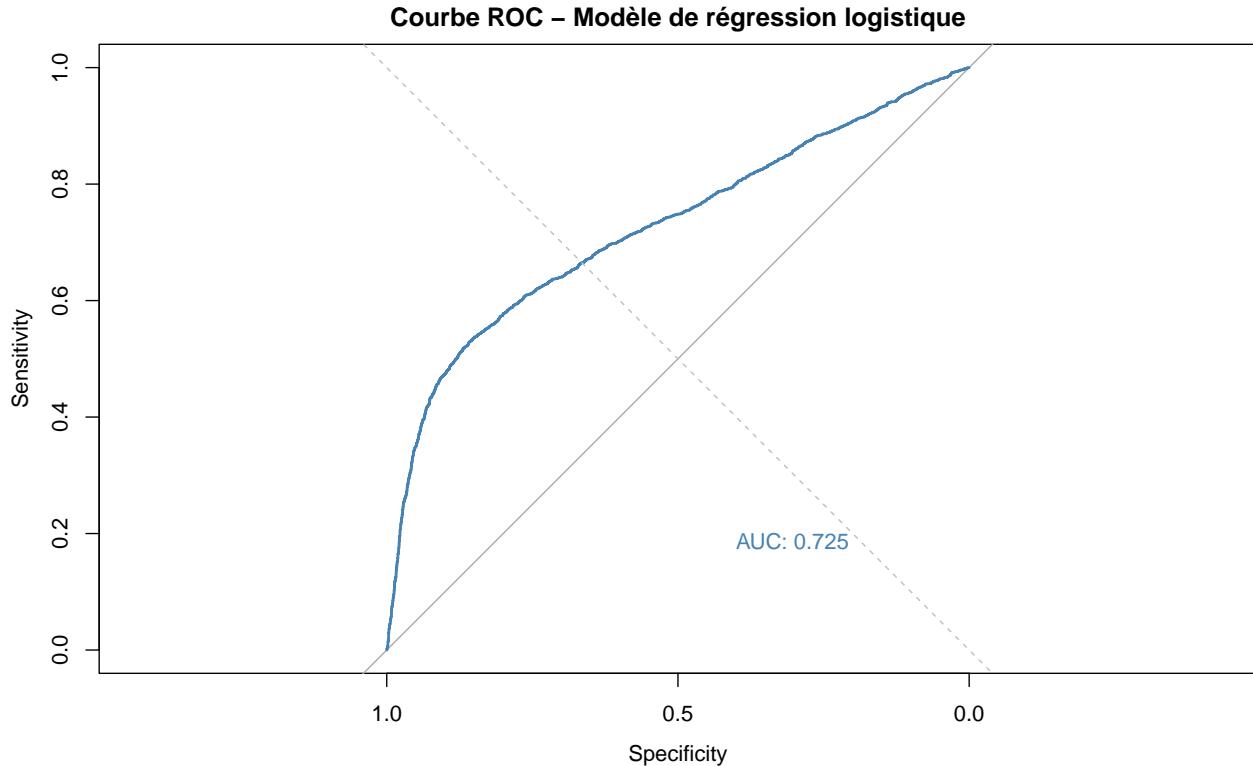
Table 9: Métriques de performance du modèle

Métrique	Valeur
Exactitude (Accuracy)	81.09%
Sensibilité (Recall)	24.57%
Spécificité	97.22%
Précision	71.57%
F1-Score	0.3659

Analyse des métriques :

- **Exactitude** : Élevée (environ 81%), mais ce chiffre est gonflé par le déséquilibre des classes.
- **Sensibilité** : Faible (environ 25%), le modèle ne détecte qu'un quart des vrais défauts. C'est une limitation importante pour la gestion du risque de crédit.
- **Spécificité** : Très élevée (environ 97%), le modèle identifie très bien les non-défauts.
- **F1-Score** : Modéré (environ 0.37), reflétant le compromis entre précision et sensibilité.

4.7.3 Courbe ROC



```
## Aire sous la courbe ROC (AUC): 0.7255
```

Analyse de la courbe ROC :

La courbe ROC représente le taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (1 - spécificité) pour différents seuils de classification.

- La courbe est nettement au-dessus de la diagonale (modèle aléatoire), indiquant un pouvoir discriminant réel.
- L'AUC d'environ 0.73 est considérée comme "acceptable" en pratique.
- Un AUC de 0.5 correspondrait à un modèle aléatoire, et 1.0 à un modèle parfait.

```
## Seuil optimal (maximisant l'indice de Youden): 0.2731
```

```
## Sensibilité au seuil optimal: 52.95 %
```

```
## Spécificité au seuil optimal: 85.83 %
```

5 ANOVA : Comparaison des Groupes

5.1 Présentation de la méthode

L'Analyse de la Variance (ANOVA) est une méthode statistique qui permet de comparer les moyennes de plusieurs groupes. Elle teste l'hypothèse nulle selon laquelle toutes les moyennes des groupes sont égales contre l'hypothèse alternative qu'au moins une moyenne diffère.

Principe mathématique : L'ANOVA décompose la variance totale en deux parties : - La variance inter-groupes (due aux différences entre les moyennes des groupes) - La variance intra-groupes (due à la variabilité à l'intérieur de chaque groupe)

Le test F compare ces deux sources de variance : $F = \frac{MS_{inter}}{MS_{intra}}$

5.2 Objectif

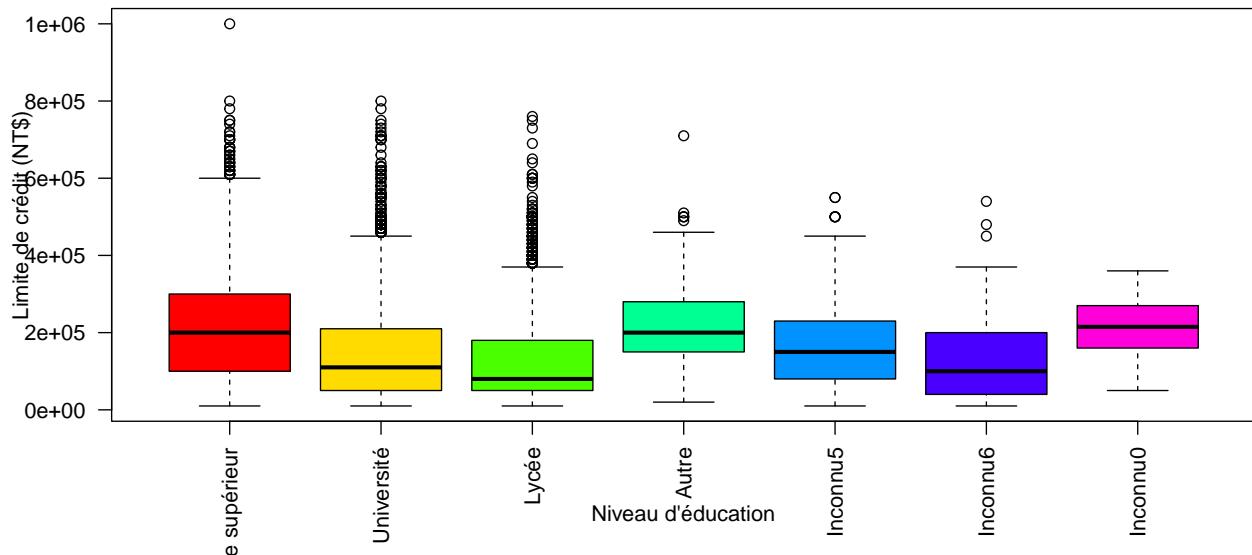
Nous utilisons l'ANOVA pour tester si la limite de crédit diffère significativement selon le niveau d'éducation des clients.

5.3 Conditions d'application

Condition	Statut	Justification
Variable dépendante quantitative	Oui	LIMIT_BAL est continue
Variable indépendante catégorielle	Oui	EDUCATION a 7 modalités
Indépendance des observations	Oui	Clients indépendants
Homogénéité des variances	À vérifier	Test de Levene ci-dessous
Normalité des résidus	Oui	n = 30 000, TCL applicable

5.4 Visualisation et résultats

Limite de crédit selon le niveau d'éducation



Analyse du boxplot :

- Les clients avec un “Diplôme supérieur” et “Autre” ont les médianes de limite de crédit les plus élevées (200 000 NT\$).
- La catégorie “Université” présente une médiane intermédiaire (110 000 NT\$).
- La catégorie “Lycée” a la médiane la plus basse (80 000 NT\$).
- Toutes les distributions sont asymétriques à droite avec de nombreuses valeurs aberrantes.
- La variabilité semble différer entre les groupes (boîtes de tailles différentes).

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df    F value    Pr(>F)
## group      6  57.706 < 2.2e-16 ***
##             29993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interprétation du test de Levene : La p-value < 0.05 indique que l'hypothèse d'homogénéité des variances est rejetée. Cependant, avec un échantillon de 30 000 observations, l'ANOVA reste robuste face à cette violation grâce au théorème central limite.

Note sur le test de Shapiro-Wilk : Pour les grands échantillons ($n > 5\ 000$), ce test est extrêmement sensible et rejette presque systématiquement l'hypothèse de normalité, même pour des écarts mineurs. L'évaluation visuelle par QQ-plot est donc plus appropriée dans ce contexte.

```
##           Df    Sum Sq   Mean Sq F value Pr(>F)
## EDUCATION      6 3.638e+13 6.063e+12   388.1 <2e-16 ***
## Residuals  29993 4.686e+14 1.562e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interprétation de l'ANOVA : La p-value extrêmement faible (< 2e-16) indique que le niveau d'éducation a un effet hautement significatif sur la limite de crédit. Les personnes avec un niveau d'éducation supérieur ont tendance à obtenir des limites de crédit plus élevées.

5.5 Test du Chi-deux : Sexe et défaut

Table 11: Tableau de contingence : Sexe vs Défaut

	Non	Oui
Homme	9015	2873
Femme	14349	3763

```
##  
## Pearson's Chi-squared test with Yates' continuity correction  
##  
## data: tab_sexe  
## X-squared = 47.709, df = 1, p-value = 4.945e-12
```

Analyse du tableau de contingence :

- Hommes : 2 873 défauts sur 11 888 (24.2%)
- Femmes : 3 763 défauts sur 18 112 (20.8%)

Le test du Chi-deux confirme une association significative ($p < 0.05$) entre le sexe et le défaut. Les hommes ont un taux de défaut légèrement mais significativement plus élevé que les femmes.

6 Régression Linéaire Simple

6.1 Présentation de la méthode

La régression linéaire simple modélise la relation entre une variable dépendante quantitative (Y) et une variable explicative quantitative (X) par une droite : $Y = \beta_0 + \beta_1 X + \epsilon$

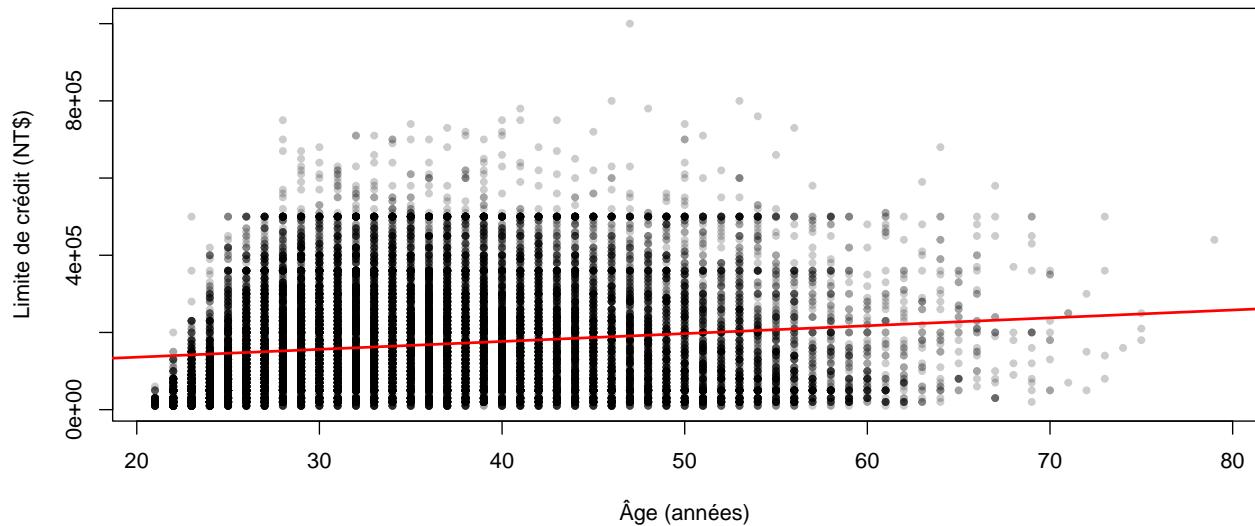
6.2 Objectif

Nous analysons la relation entre l'âge des clients et leur limite de crédit.

6.3 Conditions d'application

Condition	Statut	À vérifier
Variables quantitatives	Oui	AGE et LIMIT_BAL continues
Relation linéaire	À vérifier	Nuage de points
Indépendance	Oui	Observations indépendantes
Homoscédasticité	À vérifier	Graphique des résidus
Normalité des résidus	À vérifier	QQ-plot

Relation entre l'âge et la limite de crédit



Analyse du nuage de points :

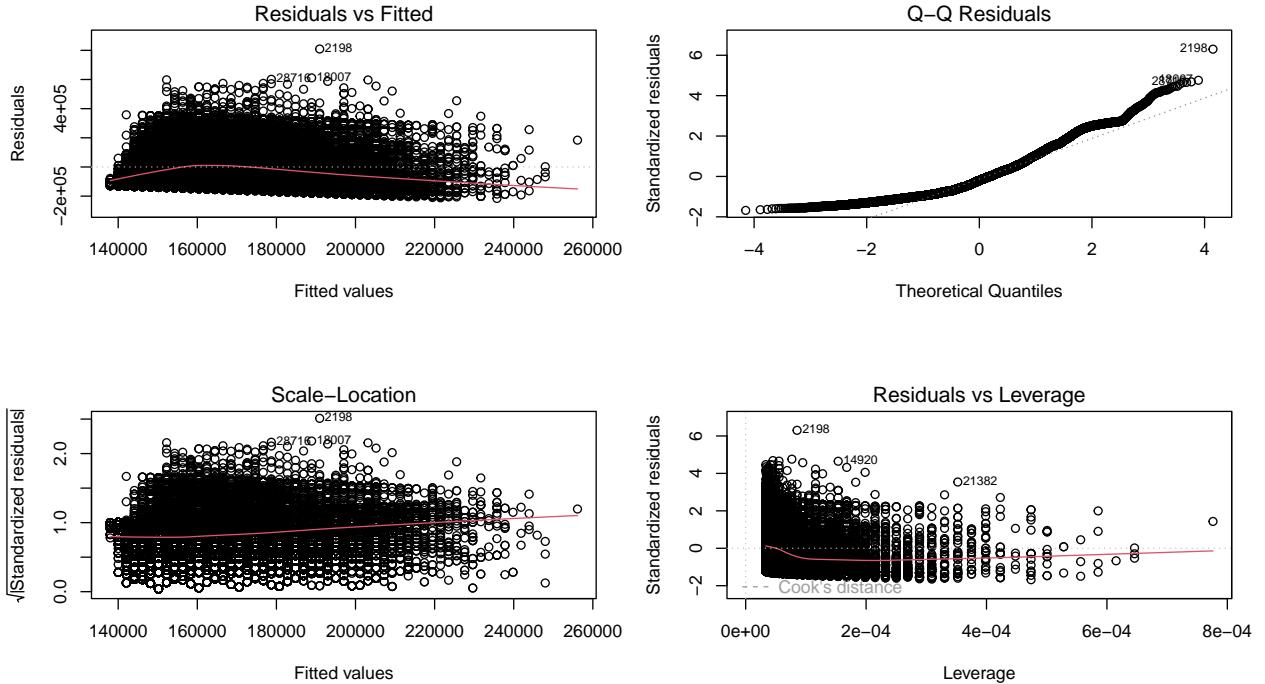
- Le nuage est très dispersé, sans tendance linéaire claire visible.
- La droite de régression (en rouge) a une pente très faible.
- La variabilité de LIMIT_BAL semble augmenter avec l'âge (forme en entonnoir), suggérant une hétéroscélasticité.

```
##
## Call:
## lm(formula = LIMIT_BAL ~ AGE, data = donnees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -215751 -102237  -25197   71098  809062
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 95203.14    2948.19   32.29 <2e-16 ***
## AGE         2036.92     80.41   25.33 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128400 on 29998 degrees of freedom
## Multiple R-squared:  0.02094, Adjusted R-squared:  0.02091
## F-statistic: 641.6 on 1 and 29998 DF,  p-value: < 2.2e-16
##
## Coefficient de détermination R2: 0.0209
##
## L'âge explique 2.09 % de la variance de la limite de crédit.
```

Interprétation des résultats :

- Le R² est extrêmement faible (environ 2%), indiquant que l'âge n'explique presque rien de la variance de la limite de crédit.

- Bien que le coefficient soit statistiquement significatif ($p < 0.05$), cela est uniquement dû à la grande taille de l'échantillon.
- En pratique, cette relation n'a **aucune valeur prédictive**.



Analyse des graphiques diagnostiques :

Graphique	Observation	Respectée ?	Commentaire
Résidus vs Fitted	Structure en entonnoir visible	Non	Hétéroscédasticité
QQ-plot	Écarts importants aux extrémités	Non	Non-normalité
Scale-Location	Tendance croissante	Non	Variance non constante
Residuals vs Leverage	Quelques points influents	Oui	Acceptable

Note : Pour les grands échantillons, le test de Shapiro-Wilk est très sensible. L'évaluation visuelle par QQ-plot est préférée.

Bilan : Trois conditions sur quatre ne sont pas respectées. La régression linéaire simple **n'est pas adaptée** pour modéliser cette relation.

7 Régression Linéaire Multiple

7.1 Présentation de la méthode

La régression linéaire multiple étend la régression simple en incluant plusieurs variables explicatives : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$

7.2 Objectif

Nous cherchons à expliquer la limite de crédit en fonction de plusieurs caractéristiques des clients.

```

## Analysis of Variance Table
##
## Response: LIMIT_BAL
##             Df      Sum Sq   Mean Sq F value    Pr(>F)
## AGE          1 1.0576e+13 1.0576e+13 799.341 < 2.2e-16 ***
## SEX          1 7.3168e+11 7.3168e+11  55.301 1.062e-13 ***
## EDUCATION    6 4.4089e+13 7.3481e+12 555.379 < 2.2e-16 ***
## MARRIAGE     3 3.6834e+12 1.2278e+12  92.799 < 2.2e-16 ***
## BILL_AMT1    1 3.9530e+13 3.9530e+13 2987.696 < 2.2e-16 ***
## PAY_AMT1     1 9.6673e+12 9.6673e+12  730.663 < 2.2e-16 ***
## Residuals 29986 3.9674e+14 1.3231e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Comparaison des AIC:

## Modèle complet - AIC: 783510.7

## Modèle sélectionné - AIC: 783510.7

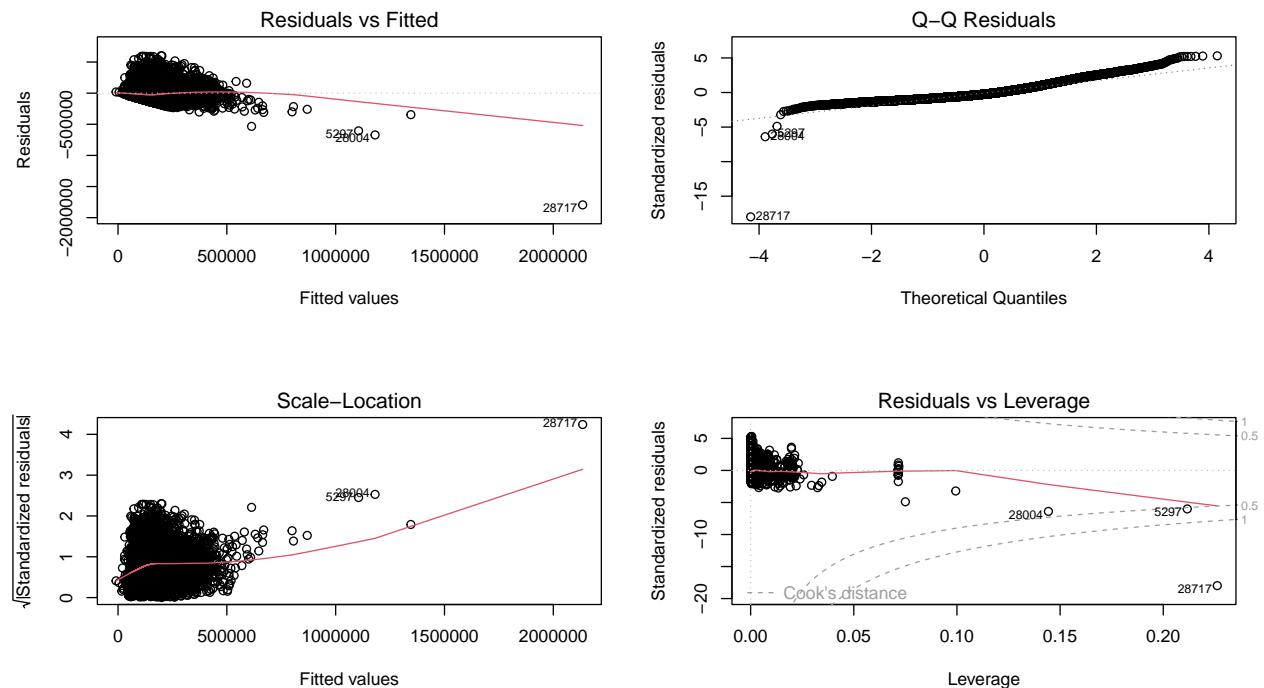
##
## Call:
## lm(formula = LIMIT_BAL ~ AGE + SEX + EDUCATION + MARRIAGE + BILL_AMT1 +
##     BILL_AMT2 + BILL_AMT3 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3, data = donnees)
##
## Residuals:
##      Min       1Q     Median       3Q      Max
## -1795062 -80091 -29369  59355  601851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.136e+05 3.753e+03 30.273 < 2e-16 ***
## AGE         2.032e+03 8.288e+01 24.521 < 2e-16 ***
## SEXFemme    1.308e+04 1.351e+03  9.682 < 2e-16 ***
## EDUCATIONUniversité -6.749e+04 1.480e+03 -45.614 < 2e-16 ***
## EDUCATIONLycée -9.729e+04 2.024e+03 -48.069 < 2e-16 ***
## EDUCATIONAutre  2.899e+03 1.029e+04   0.282  0.7782
## EDUCATIONInconnu5 -6.674e+04 6.882e+03  -9.698 < 2e-16 ***
## EDUCATIONInconnu6 -1.020e+05 1.595e+04  -6.395 1.63e-10 ***
## EDUCATIONInconnu0  9.198e+03 3.035e+04   0.303  0.7619
## MARRIAGECélibataire -1.892e+04 1.510e+03 -12.531 < 2e-16 ***
## MARRIAGEAutre    -7.667e+04 6.397e+03 -11.984 < 2e-16 ***
## MARRIAGEInconnu   9.786e+02 1.552e+04   0.063  0.9497
## BILL_AMT1        6.209e-01 3.265e-02 19.014 < 2e-16 ***
## BILL_AMT2        -2.863e-01 4.648e-02 -6.159 7.40e-10 ***
## BILL_AMT3         8.054e-02 3.367e-02   2.392  0.0168 *
## PAY_AMT1         8.962e-01 5.025e-02 17.835 < 2e-16 ***
## PAY_AMT2         3.999e-01 4.016e-02   9.959 < 2e-16 ***
## PAY_AMT3         7.781e-01 4.032e-02 19.295 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```

```

## Residual standard error: 113500 on 29982 degrees of freedom
## Multiple R-squared:  0.2357, Adjusted R-squared:  0.2353
## F-statistic: 543.9 on 17 and 29982 DF, p-value: < 2.2e-16

```



Analyse des graphiques diagnostiques :

- **Résidus vs Fitted** : Structure en entonnoir marquée → hétéroscédasticité
- **QQ-plot** : Écarts importants dans les queues → non-normalité
- **Scale-Location** : Tendance croissante → variance non constante

Table 14: VIF - Régression multiple

	GVIF	Df	$GVIF^{(1/(2*Df))}$
AGE	1.36	1	1.17
SEX	1.02	1	1.01
EDUCATION	1.10	6	1.01
MARRIAGE	1.32	3	1.05
BILL_AMT1	13.47	1	3.67
BILL_AMT2	25.50	1	5.05
BILL_AMT3	12.70	1	3.56
PAY_AMT1	1.61	1	1.27
PAY_AMT2	1.99	1	1.41
PAY_AMT3	1.17	1	1.08

Table 15: Coefficients avec IC à 95%

	Coefficient	IC_inf	IC_sup
(Intercept)	113625.73	106268.88	120982.59

	Coefficient	IC_inf	IC_sup
AGE	2032.39	1869.93	2194.84
SEXFemme	13079.23	10431.55	15726.91
EDUCATIONUniversité	-67492.43	-70392.60	-64592.26
EDUCATIONLycée	-97291.12	-101258.26	-93323.98
EDUCATIONAutre	2899.38	-17277.83	23076.59
EDUCATIONInconnu5	-66744.56	-80233.55	-53255.56
EDUCATIONInconnu6	-101969.37	-133223.41	-70715.33
EDUCATIONInconnu0	9197.59	-50292.12	68687.31
MARRIAGECélibataire	-18916.87	-21875.73	-15958.01
MARRIAGEAutre	-76667.71	-89206.75	-64128.66
MARRIAGEInconnu	978.59	-29436.81	31393.99
BILL_AMT1	0.62	0.56	0.68
BILL_AMT2	-0.29	-0.38	-0.20
BILL_AMT3	0.08	0.01	0.15
PAY_AMT1	0.90	0.80	0.99
PAY_AMT2	0.40	0.32	0.48
PAY_AMT3	0.78	0.70	0.86

```
##  
## R² : 0.2357
```

```
## R² ajusté: 0.2353
```

Bilan des conditions d'application :

Condition	Respectée ?	Commentaire
Variable dépendante quantitative	Oui	LIMIT_BAL continue
Indépendance	Oui	Observations indépendantes
Absence de multicolinéarité	Partielle	VIF élevés pour BILL_AMT
Homoscédasticité	Non	Entonnoir visible
Normalité des résidus	Non	QQ-plot déviant

Note : Pour les grands échantillons, le test de Shapiro-Wilk est très sensible. L'évaluation visuelle par QQ-plot est préférée.

Conclusion : Le R² reste modeste malgré l'ajout de variables. Les hypothèses ne sont pas respectées, limitant la validité des inférences. Cette méthode n'est pas optimale pour ce jeu de données mais illustre l'importance de la vérification des conditions.

8 Conclusion

8.1 Résumé des résultats

Méthode	Résultat principal	Validité
Analyse descriptive	22% de défauts, déséquilibre des classes	Valide
ACP	Axes : endettement + remboursement	Valide
Régression logistique	AUC acceptable, PAY_0 prédicteur clé	Valide

Méthode	Résultat principal	Validité
ANOVA	Éducation influence LIMIT_BAL	Valide
Régression linéaire simple	R ² très faible, non adaptée	Non valide
Régression linéaire multiple	R ² modeste, hypothèses non respectées	Partielle

8.2 Limites de l'étude

- Multicolinéarité entre variables BILL_AMT
- Déséquilibre des classes affectant les prédictions (sensibilité limitée à 25%)
- Régressions linéaires non adaptées (hypothèses violées)
- Données de Taïwan, généralisation limitée

9 Références

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473-2480.

10 Annexes

Annexe A : Code R - Chargement et Préparation des Données

```
# Installation des packages si nécessaire
packages_requis <- c("readxl", "ade4", "factoextra", "DescTools",
                     "GGally", "ggplot2", "corrplot", "pROC", "car", "knitr")

for (pkg in packages_requis) {
  if (!require(pkg, character.only = TRUE, quietly = TRUE)) {
    install.packages(pkg, repos = "https://cloud.r-project.org")
    library(pkg, character.only = TRUE)
  }
}

# Chargement des données
donnees <- read_excel("default of credit card clients.xls", skip = 1)

# Renommer les variables
colnames(donnees) <- c("ID", "LIMIT_BAL", "SEX", "EDUCATION", "MARRIAGE", "AGE",
                       "PAY_0", "PAY_2", "PAY_3", "PAY_4", "PAY_5", "PAY_6",
                       "BILL_AMT1", "BILL_AMT2", "BILL_AMT3", "BILL_AMT4",
                       "BILL_AMT5", "BILL_AMT6",
                       "PAY_AMT1", "PAY_AMT2", "PAY_AMT3", "PAY_AMT4",
                       "PAY_AMT5", "PAY_AMT6",
                       "DEFAULT")

# Suppression de l'ID
donnees <- donnees[, -1]

# Conversion des variables catégorielles
donnees$SEX <- factor(donnees$SEX, levels = c(1, 2), labels = c("Homme", "Femme"))
donnees$EDUCATION <- factor(donnees$EDUCATION,
                            levels = c(1, 2, 3, 4, 5, 6, 0),
                            labels = c("Diplôme supérieur", "Université", "Lycée",
                                      "Autre", "Inconnu5", "Inconnu6", "Inconnu0"))
donnees$MARRIAGE <- factor(donnees$MARRIAGE,
                            levels = c(1, 2, 3, 0),
                            labels = c("Marié", "Célibataire", "Autre", "Inconnu"))
donnees$DEFAULT <- factor(donnees$DEFAULT, levels = c(0, 1), labels = c("Non", "Oui"))
donnees$DEFAULT_NUM <- as.numeric(donnees$DEFAULT) - 1
```

Annexe B : Code R - Analyse Descriptive

```
# Tableau de fréquences
tab_default <- data.frame(
  Statut = c("Pas de défaut", "Défaut"),
  Effectif = as.vector(table(donnees$DEFAULT)),
  Pourcentage = as.vector(round(prop.table(table(donnees$DEFAULT)) * 100, 2))
)
kable(tab_default, caption = "Distribution du défaut de paiement")
```

```

# Graphique
ggplot(donnees, aes(x = DEFAULT, fill = DEFAULT)) +
  geom_bar() +
  scale_fill_manual(values = c("steelblue", "coral")) +
  labs(title = "Distribution du défaut de paiement",
       x = "Défaut de paiement", y = "Effectif") +
  theme_minimal() +
  theme(legend.position = "none")
var_quant <- c("LIMIT_BAL", "AGE", "BILL_AMT1", "PAY_AMT1")

stats_desc <- data.frame(
  Variable = var_quant,
  Moyenne = sapply(donnees[var_quant], mean),
  Écart_type = sapply(donnees[var_quant], sd),
  Médiane = sapply(donnees[var_quant], median),
  Min = sapply(donnees[var_quant], min),
  Max = sapply(donnees[var_quant], max)
)
kable(stats_desc, digits = 2, caption = "Statistiques descriptives des variables quantitatives principales")

# Distribution par sexe et défaut
tab_sexe <- prop.table(table(donnees$SEX, donnees$DEFAULT), margin = 1) * 100

ggplot(donnees, aes(x = SEX, fill = DEFAULT)) +
  geom_bar(position = "fill") +
  scale_fill_manual(values = c("steelblue", "coral"), name = "Défaut") +
  labs(title = "Proportion de défaut selon le sexe",
       x = "Sexe", y = "Proportion") +
  theme_minimal()
par(mfrow = c(2, 2))

boxplot(LIMIT_BAL ~ DEFAULT, data = donnees,
        main = "Limite de crédit selon le défaut",
        xlab = "Défaut", ylab = "Limite de crédit (NT$)",
        col = c("steelblue", "coral"))

boxplot(AGE ~ DEFAULT, data = donnees,
        main = "Âge selon le défaut",
        xlab = "Défaut", ylab = "Âge (années)",
        col = c("steelblue", "coral"))

boxplot(BILL_AMT1 ~ DEFAULT, data = donnees,
        main = "Montant facture (Sept) selon le défaut",
        xlab = "Défaut", ylab = "Montant (NT$)",
        col = c("steelblue", "coral"))

boxplot(PAY_AMT1 ~ DEFAULT, data = donnees,
        main = "Montant paiement (Sept) selon le défaut",
        xlab = "Défaut", ylab = "Montant (NT$)",
        col = c("steelblue", "coral"))

var_corr <- c("LIMIT_BAL", "AGE", "BILL_AMT1", "BILL_AMT2", "BILL_AMT3",
            "PAY_AMT1", "PAY_AMT2", "PAY_AMT3")
mat_cor <- cor(donnees[var_corr], use = "complete.obs")

```

```

corrplot(mat_cor, method = "color", type = "upper",
         tl.cex = 0.8, tl.col = "black",
         addCoef.col = "black", number.cex = 0.6,
         title = "Matrice de corrélation", mar = c(0, 0, 2, 0))

```

Annexe C : Code R - Analyse en Composantes Principales

```

# Sélection des variables pour l'ACP
var_acp <- c("LIMIT_BAL", "AGE",
            "BILL_AMT1", "BILL_AMT2", "BILL_AMT3", "BILL_AMT4",
            "BILL_AMT5", "BILL_AMT6",
            "PAY_AMT1", "PAY_AMT2", "PAY_AMT3", "PAY_AMT4",
            "PAY_AMT5", "PAY_AMT6")

donnees_acp <- donnees[, var_acp]

# Réalisation de l'ACP
ACP <- dudi.pca(donnees_acp, scannf = FALSE, nf = 5)
# Valeurs propres et variance expliquée
poids <- ACP$eig / sum(ACP$eig)
var_cumul <- cumsum(poids)

tab_vp <- data.frame(
  Composante = 1:length(ACP$eig),
  Valeur_propre = round(ACP$eig, 3),
  Variance_expliquée = round(poids * 100, 2),
  Variance_cumulée = round(var_cumul * 100, 2)
)
kable(tab_vp, caption = "Valeurs propres et variance expliquée")

# Graphique des éboulis
fviz_eig(ACP, addlabels = TRUE,
          main = "Graphique des éboulis",
          xlab = "Composantes principales",
          ylab = "Pourcentage de variance expliquée")
fviz_pca_var(ACP, axes = c(1, 2),
              repel = TRUE,
              col.var = "contrib",
              gradient.cols = c("blue", "yellow", "red"),
              title = "Cercle des corrélations - Axes 1 et 2")
var_res <- get_pca_var(ACP)
kable(round(var_res$coord[, 1:3], 3),
      caption = "Coordonnées des variables sur les 3 premiers axes")
# Échantillon pour la visualisation
set.seed(123)
idx <- sample(1:nrow(donnees), 3000)

fviz_pca_ind(ACP,
             geom = "point",
             col.ind = donnees$DEFAULT,
             palette = c("steelblue", "coral"),

```

```

    addEllipses = TRUE,
    ellipse.type = "convex",
    legend.title = "Défaut",
    title = "Projection des individus - Axes 1 et 2",
    select.ind = list(ind = idx))

```

Annexe D : Code R - Régression Logistique

```

set.seed(42)
n <- nrow(donnees)
indices_train <- sample(1:n, size = 0.7 * n)
donnees_train <- donnees[indices_train, ]
donnees_test <- donnees[-indices_train, ]

cat("Taille de l'ensemble d'entraînement:", nrow(donnees_train), "\n")
cat("Taille de l'ensemble de test:", nrow(donnees_test), "\n")
modele_complet <- glm(DEFAULT_NUM ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE +
                         PAY_0 + PAY_2 + PAY_3 + PAY_4 + PAY_5 + PAY_6 +
                         BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT4 +
                         BILL_AMT5 + BILL_AMT6 +
                         PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 +
                         PAY_AMT5 + PAY_AMT6,
                         family = binomial(link = "logit"),
                         data = donnees_train)
# Sélection stepwise basée sur l'AIC
modele_step <- step(modele_complet, direction = "both", trace = 0)

cat("AIC du modèle complet:", AIC(modele_complet), "\n")
cat("AIC du modèle réduit:", AIC(modele_step), "\n")
# Calcul des Odds Ratios
OR <- exp(coef(modele_step))
IC_OR <- exp(confint(modele_step))

tableau_OR <- data.frame(
  Variable = names(OR),
  OR = round(OR, 4),
  IC_inf = round(IC_OR[, 1], 4),
  IC_sup = round(IC_OR[, 2], 4)
)
kable(tableau_OR, caption = "Odds Ratios avec intervalles de confiance à 95%", row.names = FALSE)
# Calcul des VIF
vif_values <- vif(modele_step)

if (is.matrix(vif_values)) {
  kable(round(vif_values, 2), caption = "Facteurs d'inflation de la variance (GVIF)")
} else {
  kable(data.frame(Variable = names(vif_values), VIF = round(vif_values, 2)),
        caption = "Facteurs d'inflation de la variance (VIF)")
}
par(mfrow = c(2, 2))

residus_dev <- residuals(modele_step, type = "deviance")

```

```

plot(fitted(modele_step), residus_dev,
      xlab = "Probabilités prédictes",
      ylab = "Résidus de déviance",
      main = "Résidus vs Probabilités prédictes",
      pch = 20, col = rgb(0, 0, 0, 0.2))
abline(h = 0, col = "red", lwd = 2)

hist(residus_dev, breaks = 50,
      main = "Distribution des résidus",
      xlab = "Résidus de déviance",
      col = "steelblue")

qqnorm(residus_dev, main = "QQ-plot des résidus")
qqline(residus_dev, col = "red", lwd = 2)

plot(modele_step, which = 4, main = "Distance de Cook")
prob_pred <- predict(modele_step, newdata = donnees_test, type = "response")
classe_pred <- ifelse(prob_pred > 0.5, 1, 0)

# On force les deux niveaux pour éviter les erreurs d'indexation
matrice_conf <- table(
  Observé = factor(donnees_test$DEFAULT_NUM, levels = c(0, 1)),
  Prédit = factor(classe_pred, levels = c(0, 1))
)
kable(matrice_conf, caption = "Matrice de confusion (seuil = 0.5)")
VP <- matrice_conf[2, 2]
VN <- matrice_conf[1, 1]
FP <- matrice_conf[1, 2]
FN <- matrice_conf[2, 1]

accuracy <- (VP + VN) / sum(matrice_conf)
sensibilite <- VP / (VP + FN)
specificite <- VN / (VN + FP)
precision <- VP / (VP + FP)
F1_score <- 2 * precision * sensibilite / (precision + sensibilite)

metriques <- data.frame(
  Métrique = c("Exactitude (Accuracy)", "Sensibilité (Recall)",
              "Spécificité", "Précision", "F1-Score"),
  Valeur = c(paste0(round(accuracy * 100, 2), "%"),
             paste0(round(sensibilite * 100, 2), "%"),
             paste0(round(specificite * 100, 2), "%"),
             paste0(round(precision * 100, 2), "%"),
             round(F1_score, 4)))
)
kable(metriques, caption = "Métriques de performance du modèle")
roc_obj <- roc(donnees_test$DEFAULT_NUM, prob_pred)

plot(roc_obj,
      main = "Courbe ROC - Modèle de régression logistique",
      col = "steelblue", lwd = 2,
      print.auc = TRUE, print.auc.x = 0.4, print.auc.y = 0.2)
abline(a = 0, b = 1, lty = 2, col = "gray")

```

```

cat("Aire sous la courbe ROC (AUC):", round(auc(roc_obj), 4), "\n")
coords_roc <- coords(roc_obj, "best", ret = c("threshold", "sensitivity", "specificity"))

cat("Seuil optimal (maximisant l'indice de Youden):", round(coords_roc$threshold, 4), "\n")
cat("Sensibilité au seuil optimal:", round(coords_roc$sensitivity * 100, 2), "%\n")
cat("Spécificité au seuil optimal:", round(coords_roc$specificity * 100, 2), "%\n")

```

Annexe E : Code R - ANOVA et Test du Chi-deux

```

boxplot(LIMIT_BAL ~ EDUCATION, data = donnees,
        main = "Limite de crédit selon le niveau d'éducation",
        xlab = "Niveau d'éducation",
        ylab = "Limite de crédit (NT$)",
        col = rainbow(7), las = 2)

anova_edu <- aov(LIMIT_BAL ~ EDUCATION, data = donnees)
# Test de Levene pour l'homogénéité des variances
levene_test <- car::leveneTest(LIMIT_BAL ~ EDUCATION, data = donnees)
print(levene_test)
summary(anova_edu)
tab_sexe <- table(donnees$SEX, donnees$DEFAULT)
kable(tab_sexe, caption = "Tableau de contingence : Sexe vs Défaut")

chi2_sexe <- chisq.test(tab_sexe)
print(chi2_sexe)

```

Annexe F : Code R - Régression Linéaire Simple

```

plot(donnees$AGE, donnees$LIMIT_BAL,
      main = "Relation entre l'âge et la limite de crédit",
      xlab = "Âge (années)",
      ylab = "Limite de crédit (NT$)",
      pch = 20, col = rgb(0, 0, 0, 0.2))

reg_simple <- lm(LIMIT_BAL ~ AGE, data = donnees)
abline(reg_simple, col = "red", lwd = 2)
summary(reg_simple)

r2_simple <- summary(reg_simple)$r.squared
cat("\nCoefficient de détermination R2:", round(r2_simple, 4), "\n")
cat("L'âge explique", round(r2_simple * 100, 2), "% de la variance de la limite de crédit.\n")
par(mfrow = c(2, 2))
plot(reg_simple)

```

Annexe G : Code R - Régression Linéaire Multiple

```

# On utilise directement les facteurs, R crée automatiquement les variables indicatrices
reg_multiple <- lm(LIMIT_BAL ~ AGE + SEX + EDUCATION + MARRIAGE + BILL_AMT1 + PAY_AMT1,
                    data = donnees)
anova(reg_multiple)
# On utilise directement les facteurs pour les variables catégorielles
reg_complet <- lm(LIMIT_BAL ~ AGE + SEX + EDUCATION + MARRIAGE +
                    BILL_AMT1 + BILL_AMT2 + BILL_AMT3 +
                    PAY_AMT1 + PAY_AMT2 + PAY_AMT3,
                    data = donnees)

reg_step_lm <- step(reg_complet, direction = "both", trace = 0)

cat("\nComparaison des AIC:\n")
cat("Modèle complet - AIC:", AIC(reg_complet), "\n")
cat("Modèle sélectionné - AIC:", AIC(reg_step_lm), "\n")
summary(reg_step_lm)
par(mfrow = c(2, 2))
plot(reg_step_lm)
vif_reg_mult <- vif(reg_step_lm)

if (is.matrix(vif_reg_mult)) {
  kable(round(vif_reg_mult, 2), caption = "VIF - Régression multiple")
} else {
  kable(data.frame(Variable = names(vif_reg_mult), VIF = round(vif_reg_mult, 2)),
        caption = "VIF - Régression multiple")
}
coef_tab <- data.frame(
  Coefficient = round(coef(reg_step_lm), 2),
  IC_inf = round(confint(reg_step_lm)[, 1], 2),
  IC_sup = round(confint(reg_step_lm)[, 2], 2)
)
kable(coef_tab, caption = "Coefficients avec IC à 95%")

r2_mult <- summary(reg_step_lm)$r.squared
r2_adj <- summary(reg_step_lm)$adj.r.squared
cat("\nR²:", round(r2_mult, 4), "\n")
cat("R² ajusté:", round(r2_adj, 4), "\n")

```