Régression sur données non-linéaires

Martin Chevalier *(INSEE)*

Cette page comporte l'ensemble des cas pratiques du module "Régression sur données non-linéaires" des 1, 2 et 3 mars 2017. Elle a été réalisée avec Rstudio sous Rmarkdown et compilée le 2017-03-19.

Cas pratiques à partir de la base M_CONTRAN

Ces premiers cas pratiques reposent sur des exploitations de la base M_CONTRAN fournie par la Banque de France. Celle-ci comporte des informations sur les nouveaux crédits contractés sur la période (la base est en date de décembre 2014), en particulier leur montant, durée initiale, taux effectif global et catégorie d'instruments financiers.

Variable	**Description**
ident	Identifiant anonymisé
mt_crdt	Montant du crédit en milliers d'euros
duree_in	Durée initiale de fixation du crédit en mois
teg	Taux effectif global (en %)
categh	Grande catégorie du crédit
gar	Prêt garanti ou non
taux	Taux fixe ou variable (et support associé)

La base est rééchantillonnée au 100ème et limitée aux variables strictement nécessaires à ces cas pratiques. Le taux effectif global est arrondi au dixième. La documentation fournie par la Banque sur les variables est disponible ici.

Partie 1 : Estimer un modèle logistique dichotomique

Cas pratique 1.1 Régression logistique simple : Montant du crédit et taux d'intérêt variable

Dans ce premier cas pratique, on cherche à évaluer le lien entre montant total du crédit et caractère variable du taux d'intérêt à l'aide d'une régression logistique simple :

$$\mathbb{P}(taux\_variable_i|montant_i) = \beta_0 + \beta_1 montant_i + \varepsilon_i$$

a. Assignez une bibliothèque au répertoire dans lequel se situe le fichier `m_contran.sas7bdat` et copiez la table dans la `WORK`. Repérez dans la table les variables renseignant sur le montant du crédit ainsi que sur le caractère variable du taux d'intérêt et utilisez les outils de la statistique descriptive pour analyser leur distribution. Créez la variable `txvar` dichotomique qui indique si le taux d'intérêt est variable (1) ou non (0).

Proposition de solution

```
/*Assignation de la bibliothèque*/
LIBNAME ces "\\chemin\vers\le\repertoire\reseau";

/*Copie de la table dans la bibliothèque WORK*/
DATA m_contran;
  SET ces.m_contran;
RUN;

/*Les variables en question sont mt_crdt et taux (pensez bien à
utiliser le menu Affichage > Afficher les noms de colonnes
pour masquer les libellés).*/

/*Distribution de mt_crdt*/
PROC UNIVARIATE DATA = m_contran;
  VAR mt_crdt;
RUN;
/*Distribution particulièrement asymétrique (moyenne >> médiane)*/

/*Tri à plat sur taux*/
PROC FREQ DATA = m_contran;
  TABLES taux;
RUN;
/*La modalité tx_fix correspond aux crédits dont le taux est fixe,
les autres au support auquel est adossé le taux quand il est variable.
Les crédits à taux fixe sont très majoritaires (de l'ordre de 90 %).*/

/*Création de la l'indicatrice txvar*/
DATA m_contran;
  SET m_contran;
  txvar = (taux NE "tx_fix");
RUN;
PROC FREQ DATA =  m_contran;
  TABLES txvar;
RUN;
```

b. Soumettez le code :

```
PROC LOGISTIC DATA = m_contran;
  MODEL txvar = mt_crdt;
RUN;
```

Parcourez la sortie de SAS pour déterminer la modalité de la variable `txvar` qui est modélisée par défaut. Utilisez l'option `DESC` pour corriger ce problème.

Proposition de solution

FIGURE 1 – Sortie associé à la question 1.1.b

Model Information	
Data Set	WORK.M_CONTRAN
Response Variable	txvar
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1579
Number of Observations Used	1579

Response Profile		
Ordered Value	**txvar**	**Total Frequency**
1	0	1430
2	1	149

Probability modeled is txvar=0.

La modalité de la variable expliquée modélisée est indiquée **sous le troisième tableau de résultats**. Par défaut, les modalités sont prises dans l'**ordre croissant** (0 puis 1 pour la variable `txvar`) et la modalité de référence est la dernière dans cet ordre (ici 1).

Pour que la modalité de référence soit 0 (et donc la modalité modélisée 1), il suffit d'utiliser l'**option `DESC` juste après le nom de la variable expliquée** :

```
PROC LOGISTIC DATA = m_contran;
  MODEL variable (DESC) = mt_crdt;
RUN;
```

c. Repérez dans les sorties produites par SAS le tableau comportant les paramètres estimés par le modèle. Quel est le signe du coefficient associé à la variable `mt_crdt` et comment l'interprétez-vous ? Que pouvez-vous dire de sa valeur ?

Proposition de solution

Les coefficients estimés par le modèle figurent dans le **septième tableau affiché par SAS** (après les statistiques d'ajustement et les tests de nullité jointe, *cf.* partie 2 du module).

Le coefficient associé à la variable `mt_crdt` est estimé à 0,000415. Ce coefficient est

FIGURE 2 – Sortie associé à la question 1.1.b (suite)

Model Information	
Data Set	WORK.M_CONTRAN
Response Variable	txvar
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1579
Number of Observations Used	1579

Response Profile		
Ordered Value	**txvar**	**Total Frequency**
1	1	149
2	0	1430

Probabilitv modeled is txvar=1.

FIGURE 3 – Sortie associée à la question 1.1.c

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.3142	0.0891	675.1180	<.0001
mt_crdt	1	0.000415	0.000176	5.5640	0.0183

positif : à un montant du crédit plus élevé est associée une **probabilité plus élevée qu'il soit à taux variable**.

En revanche, contrairement au cas du modèle linéaire classique **il n'est pas possible d'interpréter directement la valeur de ce coefficient**. Les transformations à utiliser pour donner du sens à ces coefficients sont abordées dans la deuxième partie de la formation.

d. Rappelez l'hypothèse nulle et l'hypothèse alternative du test de significativité du coefficient β_1 associé à la variable `mt_crdt`. Quelles sont les valeurs de la statistique et de la p-valeur associées à ce test ? Menez ce test au seuil de 5 % puis de 1 %, d'abord en utilisant la p-valeur puis en comparant la valeur de la statistique de test aux quantiles

d'une loi du χ^2 à 1 degré de liberté.

Proposition de solution
Le test de significativité du coefficient β_1 se formule :

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 \neq 0$$

Remarque Ce test est celui mené par le logiciel. On pourrait cependant tout à fait mener un test unilatéral du type :

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 > 0$$

(en ajustant alors le quantile de la loi du χ^2 correspondant à la valeur critique).

Les valeurs de la statistique et de la p-valeur associées à ce test se lisent dans le tableau de l'estimation des paramètres (*cf.* question précédente). La statistique de test vaut ici 5,5640 et la p-valeur 0,0183.
En interprétant la p-valeur :
— 0,0183 < 0,05 donc on peut rejeter l'hypothèse de nullité de β_1 au seuil de 5 % ;
— 0,0183 > 0,01 donc on ne peut pas rejeter l'hypothèse de nullité de β_1 au seuil de 1 %.

En interprétant la valeur de la statistique de test :
— le quantile à 95 % d'une loi du χ^2 à 1 degré de liberté est 3,84, ainsi la statistique de test 5,5640 est supérieure à la valeur critique du test au seuil de 5 % donc on peut rejeter l'hypothèse de nullité de β_1 au seuil de 5 % ;
— le quantile à 99 % d'une loi du χ^2 à 1 degré de liberté est 6,63, ainsi la statistique de test 5,5640 n'est pas supérieure à la valeur critique du test au seuil de 1 % donc on ne peut pas rejeter l'hypothèse de nullité de β_1 au seuil de 1 %.

Remarque On peut indifféremment mener ce test en comparant la statistique de test fournie par SAS au quantile de niveau $1-\alpha$ d'une loi du χ^2 à 1 degré de liberté ou en comparant la racine carrée de la statistique de test au quantile de niveau $1-\alpha/2$ d'une loi normale centrée réduite. En effet, si la variable aléatoire X suit une loi normale centrée réduite alors par définition la variable aléatoire

$$Z = X^2$$

suit loi du χ^2 à 1 degré de liberté.
On observe ainsi que $\sqrt{5,5640} = 2,36$ et $2,36 > 1,96$ (on rejette H_0 au seuil de 5 %) mais $2,36 < 2,58$ (on ne rejette pas H_0 au seuil de 1 %). Les conclusions sont bien identiques.

e. Utilisez l'option `CLPARM` pour afficher les intervalles de confiance à 95 % des coefficients. Recalculez celui associé à la variable `mt_crdt` manuellement à partir des éléments des questions précédentes. Menez sur cette base une dernière fois le test de significativité du coefficient de la variable `mt_crdt` au seuil de 5 %.

Proposition de solution

L'option `CLPARM` s'utilise dans l'instruction `MODEL` en indiquant le type d'intervalle de confiance à construire : `WALD` construit un intervalle sous l'hypothèse que la distribution des paramètres suit une loi normale (au moins asymptotiquement).

```
PROC LOGISTIC DATA = m_contran;
  MODEL txvar (DESC) = mt_crdt / CLPARM = WALD;
RUN;
```

FIGURE 4 – Sortie associée à la question 1.1.e

Parameter Estimates and Wald Confidence Intervals			
Parameter	**Estimate**	**95% Confidence Limits**	
Intercept	-2.3142	-2.4887	-2.1396
mt_crdt	0.000415	0.000070	0.000759

Il est facile de recalculer l'intervalle de confiance à 95 % du coefficient associé à la variable `mt_crdt` à partir des éléments des tableaux précédents, en particulier l'estimation ponctuelle du paramètre ($\hat{\beta}_1$) et son erreur-standard ($\hat{\sigma}_{\beta_1}$) . En effet :

$$IC_{95\ \%}(\beta_1) = [\hat{\beta}_1 - 1,96 \times \hat{\sigma}_{\beta_1}; \hat{\beta}_1 + 1,96 \times \hat{\sigma}_{\beta_1}]$$

soit ici

$$\begin{aligned} IC_{95\ \%}(\beta_1) &= [0{,}000415 - 1,96 \times 0{,}000176; 0{,}000415 + 1,96 \times 0{,}000176] \\ &= [0,00007; 0,00076] \end{aligned}$$

La dernière manière de mener le test de nullité de β_1 est très directe à partir de l'intervalle de confiance : 0 n'appartient pas à l'intervalle de confiance de β_1 à 95 % aussi on peut rejeter l'hypothèse de nullité de β_1 au seuil de 5 %.

Cas pratique 1.2 Régression logistique multiple : Caractéristiques des crédits avec un taux d'intérêt variable

Dans ce deuxième cas pratique, on cherche à généraliser les résultats obtenus précédemment en estimant un modèle plus complet comportant la catégorie d'instrument financier (variable `categh`) et la durée initiale du crédit (variable `duree_in`).

a. Quelle est la nature de la variable `categh` ? Menez une analyse univariée de cette variable. Quelles conséquences cela a-t-il sur la manière d'intégrer cette variable à la régression logistique (comme à toute régression d'ailleurs) ?

Proposition de solution

La variable `categh` est une variable de nature qualitative (typologie d'instrument financier). À ce titre, on procède à son analyse avec une `PROC FREQ` :

```
/*Analyse de la variable categh*/
PROC FREQ DATA = m_contran;
  TABLES categh;
RUN;
```

Le fait que cette variable soit une variable qualitative à plusieurs modalités rend sa dichotomisation impérative en vue de son intégration dans le modèle. Pour chaque modalité de la variable, il convient de créer une variable indicatrice indiquant pour chaque individu si la modalité est prise (1) ou pas (0).

b. Dichotomisez manuellement cette variable et ajoutez-la au modèle précédent en prenant la modalité `tresor` comme référence (sans ajouter `duree_in` pour l'instant). Comparez vos résultats à ceux obtenus en soumettant le code :

```
PROC LOGISTIC DATA = m_contran;
  CLASS categh;
  MODEL txvar (DESC) = mt_crdt categh;
RUN;
```

D'où provient la différence selon vous et comment la corrigeriez-vous ?

Proposition de solution

```
/*Dichotomisation manuelle de la variable categh*/
DATA m_contran;
  SET m_contran;
  immobi = (categh = 'immobi');
  invest = (categh = 'invest');
  tresor = (categh = 'tresor');
RUN;

/*Ajout au modèle du cas pratique précédent*/
PROC LOGISTIC DATA = m_contran;
  MODEL txvar (DESC) = mt_crdt immobi invest;
RUN;
/*Note : on n'intègre pas la variable tresor
pour qu'elle constitue la modalité de référence
de la variable qualitative categh.
```

Les coefficients des deux modalités associées à la variable categh évoluent entre les deux estimations, ce qui ne devrait pas être le cas. La raison est à trouver dans la manière dont SAS dichotomise par défaut les variables, qui n'est pas très naturelle. Pour contraindre SAS à dichotomiser les variables comme elles l'auraient été naturellement, il suffit d'utiliser l'option PARAM = REF dans l'instruction CLASS :

FIGURE 5 – Sortie associée à la question 1.2.b - Tableau des coefficients avec dichotomisation manuelle

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.9331	0.1012	364.7230	<.0001
mt_crdt	1	0.000403	0.000178	5.1418	0.0234
immobi	1	-0.9739	0.5239	3.4561	0.0630
invest	1	-1.2084	0.2238	29.1498	<.0001

FIGURE 6 – Sortie associée à la question 1.2.b - Tableau des coefficients avec dichotomisation automatique

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.6605	0.1880	200.2363	<.0001
mt_crdt		1	0.000403	0.000178	5.1418	0.0234
categh	**immobi**	1	-0.2465	0.3509	0.4934	0.4824
categh	**invest**	1	-0.4810	0.2199	4.7833	0.0287

```
/*Estimation avec PARAM = REF*/
PROC LOGISTIC DATA = m_contran;
  CLASS categh / PARAM = REF;
  MODEL txvar (DESC) = mt_crdt categh;
RUN;
```

c. Comment interprétez-vous les coefficients associés aux deux modalités de la variable `categh` figurant dans le tableau de résultat ? Utilisez l'option `REF` de l'instruction `CLASS` pour modifier la modalité de référence et la placer sur `immobi`. Quels changements dans les résultats cela induit-il ?

Proposition de solution

Les coefficients associés aux deux modalités de la variable `categh` sont négatifs : le fait que le crédit soit un crédit immobilier ou d'investissement **plutôt qu'une avance de trésorerie** est associé à une probabilité plus faible que son taux soit variable, **à montant égal par ailleurs**. Dans le cas des investissements, le coefficient est statistiquement significatif au seuil de 1 % (p-valeur inférieure à 0,0001) ; dans le cas

FIGURE 7 – Sortie associée à la question 1.2.b - Tableau des coefficients avec dichotomisation automatique et `PARAM = REF`

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-1.9331	0.1012	364.7230	<.0001
mt_crdt		1	0.000403	0.000178	5.1418	0.0234
categh	**immobi**	1	-0.9739	0.5239	3.4561	0.0630
categh	**invest**	1	-1.2084	0.2238	29.1498	<.0001

des crédits immobiliers en revanche, il n'est significatif qu'au seuil de 10 % (p-valeur inférieure à 0,10 mais supérieure à 0,05).

```
/*Changement de la modalité de référence avec REF = */
PROC LOGISTIC DATA = m_contran;
  CLASS categh (REF = 'immobi') / PARAM = REF;
  MODEL txvar (DESC) = mt_crdt categh;
RUN;
```

FIGURE 8 – Sortie associée à la question 1.2.c

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.9070	0.5148	31.8841	<.0001
mt_crdt		1	0.000403	0.000178	5.1418	0.0234
categh	**invest**	1	-0.2345	0.5520	0.1804	0.6710
categh	**tresor**	1	0.9739	0.5239	3.4561	0.0630

Seuls les coefficients associés à la variable `categh` et la constante sont affectés (pas celui associé à la variable `mt_crdt`). Le fait de choisir la modalité "intermédiaire" de `categh` en termes de recours aux crédits à taux variable permet d'avoir dans le modèle un coefficient positif et un coefficient négatif à interpréter, les tests de significativité permettant de mieux juger des écarts entre les trois catégories d'instrument financier. En l'occurrence ici, il ressort qu'à montant égal par ailleurs, le degré de recours aux crédits à taux variables est non-significativement différent entre les crédits immobiliers et les crédits d'investissement, et supérieur pour les avances de trésorerie (mais pas significativement différent au seuil de 5 %).

d. Ajoutez la variable `duree_in` dans le modèle et repérez parmi les tableaux de résultat celui permettant de re-calculer la log-vraisemblance du modèle. Comparez-la à celle du modèle logistique simple estimé dans le cas pratique 1.1 Comment interprétez-vous cet écart ?

Proposition de solution

FIGURE 9 – Sortie associée à la question 1.2.d

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	988.934	915.885
SC	994.299	942.707
-2 Log L	986.934	905.885

Il s'agit du sixième tableau, qui porte sur les statistiques d'ajustement du modèle. En dernière ligne et deuxième colonne de ce tableau figure la valeur de $-2 \times \ln(L_n)$ (appelée parfois *deviance*). Pour obtenir la valeur de la log-vraisemblance atteinte par le modèle, il suffit de calculer :

$$\ell_2 = -\frac{905,885}{2} = -452,94$$

En menant le même exercice dans le modèle logistique simple du cas pratique 1.1, on obtient -486,48, donc le modèle complet du cas pratique 1.2 présente une vraisemblance supérieure au premier ($-452,94 > -486,48$). Ce second modèle comportant trois variables de plus (dont `categh` sous la forme de deux indicatrices), ce résultat est tout à fait attendu.

e. Utilisez l'instruction `OUTPUT` pour récupérer dans une table les probabilités prédites par le modèle. Utilisez des indicateurs de statistique descriptive pour comparer ces probabilités à la variable `txvar`. Le modèle vous semble-t-il de bonne qualité à l'aune de ces éléments ?

Proposition de solution

```
/*Récupération des prédictions dans la table m_contran_pred*/
PROC LOGISTIC DATA = m_contran;
  CLASS categh (REF = 'immobi') / PARAM = REF;
  MODEL txvar (DESC) = mt_crdt categh duree_in;
  OUTPUT OUT = m_contran_pred P = pred;
RUN;
```

```
/*Distribution des probabilités prédites selon la variable txvar*/
PROC MEANS DATA = m_contran_pred;
        VAR pred;
        CLASS txvar;
RUN;
```

FIGURE 10 – Sortie associée à la question 1.2.e

Analysis Variable : pred Estimated Probability						
txvar	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1430	1430	0.0888017	0.0648780	0.0090908	0.9982480
1	149	149	0.1477440	0.1066081	0.0079338	0.9735082

La probabilité que le crédit soit à taux variable est en moyenne sensiblement plus élevée pour les crédit effectivement à taux variable que pour les autres. Néanmoins, certains crédits à taux fixe ont une probabilité d'être à taux variable prédite par le modèle élevée (faux positifs), tandis que certains crédits à taux variable ont une probabilité prédite par le modèle faible (faux négatifs). L'analyse de la répartition de ces deux types d'**erreurs de classement** est au coeur de la construction de la **courbe ROC** (*cf.* partie 2 du module).

Partie 2 : Interpréter un modèle logistique dichotomique

Cas pratique 1.3 Indicateurs de qualité du modèle

L'objectif de ce cas pratique est de comparer la qualité d'ajustement de différentes modélisations de la probabilité que le taux d'intérêt d'un crédit soit variable. En particulier, on considère les modèles :

$$m1 \ : \ taux\_variable_i = \beta_0 + \beta_1 montant_i + \varepsilon_i$$

$$m2 \ : \ taux\_variable_i = \beta_0 + \beta_1 montant_i + \beta_2 immobi_i + \beta_3 tresor_i + \varepsilon_i$$

$$m3 \ : \ taux\_variable_i = \beta_0 + \beta_1 montant_i + \beta_2 immobi_i + \beta_3 tresor_i + \beta_4 duree\_in_i + \varepsilon_i$$

$$m4 \ : \ taux\_variable_i = \beta_0 + \beta_1 montant_i + \beta_2 immobi_i + \beta_3 tresor_i + \beta_4 duree\_in_i + \beta_5 gar_i + \varepsilon_i$$

où `gar` est la variable indiquant si le crédit fait l'objet d'un garantie ou non.

a. Construisez un tableau dans lequel vous reportez la log-vraisemblance, l'AIC et le SC associés à ces quatre modèles. Commentez l'évolution de ces trois indicateurs.

Proposition de solution

```
PROC LOGISTIC DATA = m_contran;
  MODEL txvar (DESC) = mt_crdt;
RUN;
PROC LOGISTIC DATA = m_contran;
  CLASS categh (REF = 'invest') / PARAM = REF;
  MODEL txvar (DESC) = mt_crdt categh;
RUN;
PROC LOGISTIC DATA = m_contran;
  CLASS categh (REF = 'invest') / PARAM = REF;
  MODEL txvar (DESC) = categh duree_in mt_crdt;
RUN;
PROC LOGISTIC DATA = m_contran;
  CLASS categh (REF = 'invest') gar(REF = 'gar_n') / PARAM = REF;
  MODEL txvar (DESC) = categh duree_in mt_crdt gar;
RUN;
```

Modèle	Log-vraisemblance	AIC	SC
m1	-486,482	976,96	987,693
m2	-467,943	943,88	965,343
m3	-452,943	915,885	942,707
m4	-451,082	914.163	946,350

En termes de commentaire :

— la log-vraisemblance augmente progressivement, ce qui est mécanique (comme pour le R^2) quand on augmente le nombre de variables dans des modèles emboîtés comme c'est le cas ici (le modèle m2 contient toutes les variables du modèle m1, le modèle m3 toutes les variables du modèle m2, etc.). L'ampleur des deux premiers "sauts" est bien supérieure à celle entre les modèles m3 et m4.

— l'AIC (*Akaike Information criterion*) diminue progressivement (en lien avec l'augmentation de la log-vraisemblance) et c'est le modèle m4 qui présente l'AIC le plus faible. Pour l'AIC c'est donc le modèle m4 qui réalise le meilleur équilibre entre qualité de la prédiction et parcimonie en termes de nombre de variables explicatives intégrées dans le modèle.

— le SC (*Schartz Criterion*, aussi appelé *Bayesian Information Criterion* ou BIC) diminue assez fortement jusqu'au modèle m3 mais remonte au modèle m4. Pour le SC c'est donc le modèle m3 qui réalise le meilleur équilibre entre qualité de la prédiction et parcimonie en termes de nombre de variables explicatives intégrées dans le modèle.

Remarque En règle générale (et c'est le cas ici), le SC conduit à des modèles plus parcimonieux que l'AIC.

b. (i) Dans le modèle m4, interprétez le test par le ratio de vraisemblance de nullité jointe de l'ensemble des paramètres du modèle.

(ii) Interprétez également le test de nullité jointe de tous les coefficients associés à la variable qualitative polytomique `categh` (effets de Type III). Comparez les statistiques de test et les p-valeurs associés aux autres variables du modèle dans le tableau "Effets de Type III" à celles accompagnant l'estimation des coefficients : que constatez-vous et comment l'expliquez-vous ?

Proposition de solution

FIGURE 11 – Sortie associée à la question 1.3.b

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	84.7715	5	<.0001
Score	88.8716	5	<.0001
Wald	61.2604	5	<.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
categh	2	52.6889	<.0001
duree_in	1	27.5473	<.0001
mt_crdt	1	3.0110	0.0827
gar	1	3.8475	0.0498

(i) Le premier tableau comporte trois tests de nullité jointe asymptotiquement équivalents (leurs résultats coïncident dès lors que le nombre d'observations dans le modèle est élevé - dépasse 100 par exemple). On s'intéresse au test par le ratio de vraisemblance (*Likelihood Ratio*, première ligne) dans la mesure où c'est celui qui est le plus naturel par rapport au mécanisme d'estimation par le maximum de vraisemblance :

— l'objectif de l'estimation par le maximum de vraisemblance est de maximiser la vraisemblance du modèle étant données les valeurs de la variable expliquée et des variables explicatives ;

— si au bout du compte le modèle final possède une vraisemblance très proche de celle d'un modèle "vide" (avec uniquement la constante), alors on a de bonne raison de penser que le modèle n'apprend rien sur la variable expliquée.

— Inversement, plus l'écart entre vraisemblance du modèle "vide" et vraisemblance du modèle avec ses variables explicatives est important, plus on va être suscep-

tible de rejeter l'hypothèse de nullité jointe de tous les coefficients du modèle (sauf la constante).

La statistique du test du ratio de vraisemblance correspond exactement à cette intuition :

$$LR = -2(\ell_0 - \ell_n)$$

où ℓ_0 est la log-vraisemblance du modèle avec uniquement la constante et ℓ_n la vraisemblance du modèle complet. Autrement dit : plus la statistique du test du ratio de vraisemblance est grande, plus on est fondé à rejeter l'hypothèse de nullité jointe de tous les coefficients du modèle.

(ii) Le second tableau comporte l'analyse des effets dits de Type III, dans lesquels c'est la nullité de l'ensemble des coefficients associés à une même variable qui est testée. En l'espèce, seule la variable `categh` est associée à plus d'un coefficient dans le modèle (exactement 2 dans la mesure où c'est une variable catégorielle à trois modalités). Dans ce cadre, l'analyse des effets de Type III permet de vérifier que, prise globalement, cette variable a tout à fait sa place dans le modèle (on peut clairement rejeter l'hypothèse de nullité jointe de ses coefficients au seuil de 1 % - sa p-valeur est inférieure à 0,0001).

Pour les autres variables du modèle, ce tableau est redondant avec les éléments d'inférence présents dans le tableau présentant l'estimation des paramètres du modèle. En effet, toutes les autres variables étant associées à un seul et unique coefficient, la significativité de la variable prise dans son ensemble est exactement équivalente à la significativité de l'unique coefficient auquel elle est associée. C'est la raison pour laquelle pour ces variables les statistiques de test et les p-valeurs sont identiques dans ce tableau et dans celui présentant l'estimation des paramètres du modèle.

c. Interprétez le pourcentage de concordance du modèle `m4` et construisez sa courbe ROC. Sélectionnez un point de cette courbe et interprétez son abscisse et son ordonnée en termes de spécificité et de sensibilité.

Proposition de solution

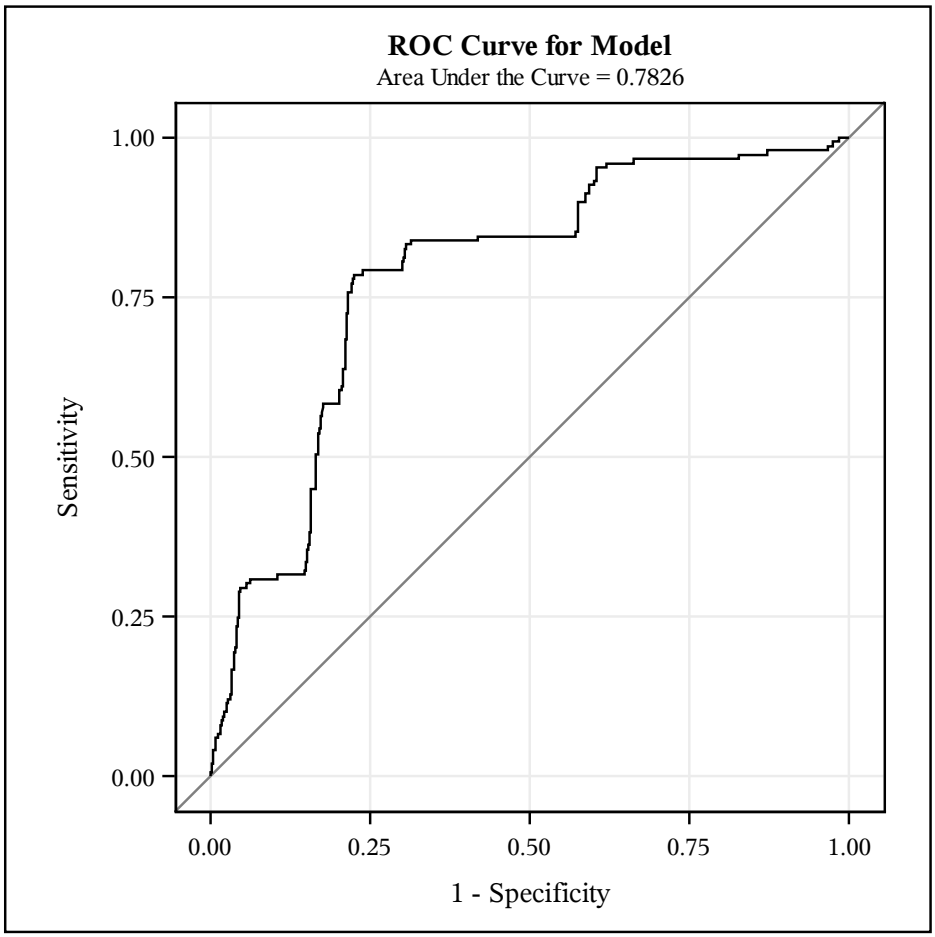
```
/*Activation des graphiques si besoin*/
ODS GRAPHICS ON;

/*Production de la courbe ROC du modèle m4*/
PROC LOGISTIC DATA = m_contran PLOTS(ONLY) = ROC;
  CLASS categh (REF = 'invest') gar(REF = 'gar_n') / PARAM = REF;
  MODEL txvar (DESC) = categh duree_in mt_crdt gar;
RUN;
```

Le pourcentage de concordance du modèle `m4` est élevé, avec plus de 78 % : dans 78 % des cas, les crédits à taux variable ont une probabilité d'être à taux variable estimée par le modèle supérieure à celle des crédits à taux fixe, et inversement.

FIGURE 12 – Sortie associée à la question 1.3.c

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.3	**Somers' D**	0.565
Percent Discordant	21.7	**Gamma**	0.565
Percent Tied	0.0	**Tau-a**	0.097
Pairs	213070	**c**	0.783

FIGURE 13 – Sortie associée à la question 1.3.c (suite)



La courbe ROC est assez erratique, en partie en raison du petit nombre d'observations utilisé ici. Avec une aire sous la courve de 0,78, une fois encore le caractère prédictif du modèle semble bon.

Si on prend par exemple le point d'abscisse 0,25 et d'ordonnée approximative 0,78, on

interprète sa position de la façon suivante : pour réussir à classer 78 % des crédits à taux variable comme tels (sensibilité), le modèle estimé doit tolérer 25 % de crédits à taux fixe classés à tort comme étant à taux variable (1 - spécificité).

d. Utilisez plusieurs instructions `ROC` dans la `PROC LOGISTIC` pour comparer les courbes ROC des quatre modèles ainsi que leurs indicateurs de qualité. Que pensez-vous de l'impact de l'ajout des différentes variables sur les propriétés **prédictives** du modèle ?

Proposition de solution

```
/*Comparaison de plusieurs courbes ROC*/
PROC LOGISTIC DATA = m_contran PLOTS(ONLY) = ROC;
    CLASS categh (REF = 'invest') gar(REF = 'gar_n') / PARAM = REF;
    m4: MODEL txvar (DESC) = mt_crdt categh duree_in gar;
    ROC 'm3' mt_crdt categh duree_in;
    ROC 'm2' mt_crdt categh;
    ROC 'm1' mt_crdt;
RUN;
```

FIGURE 14 – Sortie associée à la question 1.3.d

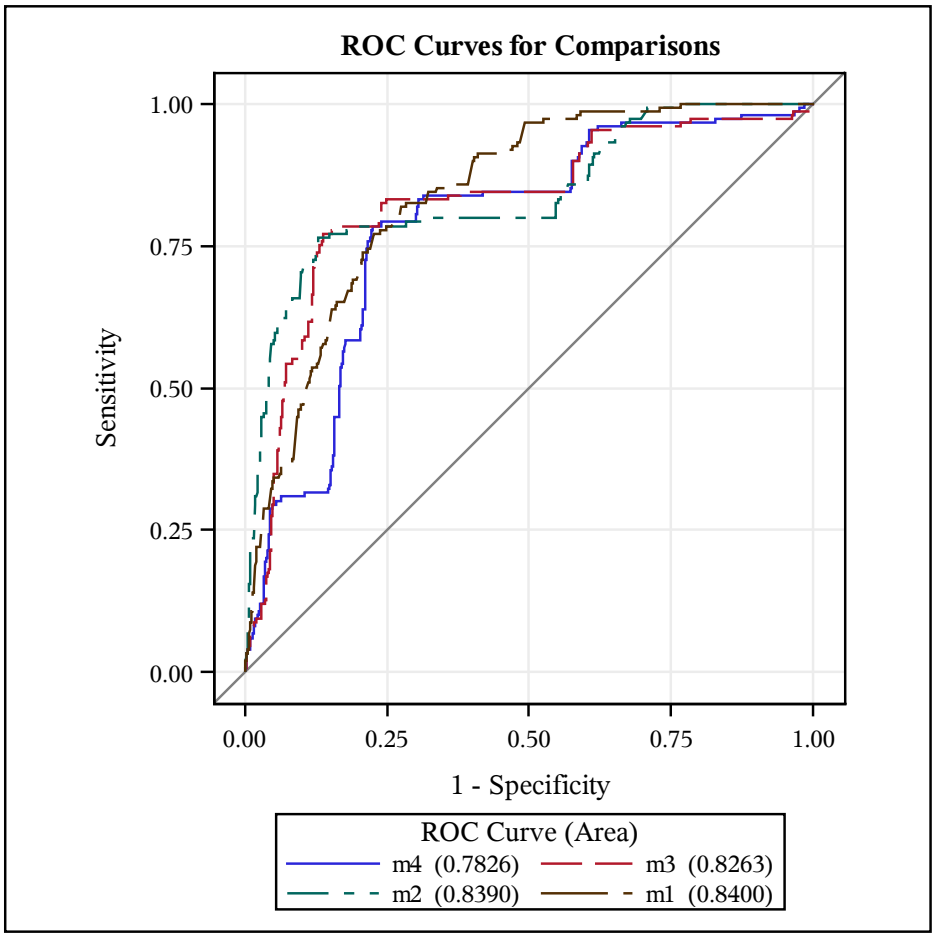


FIGURE 15 – Sortie associée à la question 1.3.d (suite)

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
m4	0.7826	0.0190	0.7455	0.8198	0.5653	0.5653	0.0967
m3	0.8263	0.0197	0.7877	0.8649	0.6526	0.6526	0.1116
m2	0.8390	0.0199	0.8000	0.8779	0.6780	0.6784	0.1159
m1	0.8400	0.0145	0.8115	0.8685	0.6799	0.6818	0.1163

Le pouvoir prédictif des modèles est dès le premier relativement fort et semble se dégrader au fur et à mesure que l'on rajoute des variables. C'est assez contre-intuitif, mais ce type de situation survient parfois :

— du point de vue de l'explication, les nouvelles variables font sens et permettent des interprétations plus riches du modèle ;

— mais du point de vue de la prédiction, elles ne modifient pas suffisamment la situation pour améliorer considérablement la qualité de celle-ci, en intruisant finalement davantage de facteurs de confusion (qui viennent légèrement détériorer la qualité de la prédiction).

À noter néanmoins que la **forme générale** de la courbe ROC évolue sensiblement selon les variables intégrées : selon le degré de spécificité souhaité (plus ou moins de crédits à taux fixes classés à tort comme étant à taux variable), choisir un modèle plutôt qu'un autre permet d'atteindre plus ou moins rapidement un niveau élevé de sensibilité.

Cas pratique 1.4 Odds-ratio et effets marginaux

a. Interprétez les *odds-ratio* associés aux modalités des variables qualitatives du modèle `m4`, à savoir `categh` et `gar`. Reformulez à partir des *odds-ratio* le test de significativité de l'association entre une variable explicative et la variable expliquée et utilisez l'intervalle de confiance à 95 des *odds-ratio* pour mener ce test au seuil de 5 %.

Proposition de solution

Si on se concentre sur les variables qualitatives (les seules pour lesquelles l'*odds-ratio* a vraiment un sens), on dira en règle générale qu'en moyenne dans l'échantillon, la probabilité que le crédit soit à taux variable est de l'ordre de 9,6 fois supérieure quand le crédit est une avance de trésorerie *plutôt qu'un investissement* et près de 2,2

FIGURE 16 – Sortie associée à la question 1.4.a

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
categh immobi vs invest	0.463	0.123	1.747
categh tresor vs invest	9.671	5.230	17.884
duree_in	1.015	1.009	1.021
mt_crdt	1.000	1.000	1.001
gar gar_y vs gar_n	1.484	1.000	2.203

fois inférieure (1/0,463 = 2,2) quand le crédit est un crédit immobilier *plutôt qu'un investissement*. Pour ce qui concerne la variable `gar`, on dira en général qu'en moyenne dans l'échantillon, la probabilité que le crédit soit à taux variable est de l'ordre de 1,5 fois supérieure quand il est garanti (*plutot que non garanti*).

Si l'on souhaite reformuler en termes d'*odds-ratio* le test de significativité de l'association entre la variable explicative et la variable expliquée (les autres variables du modèle étant égales par ailleurs), on peut repartir de l'idée que si pour le coefficient la valeur traduisant l'absence de relation entre les variables est 0, pour l'*odds-ratio* cette valeur est 1 ($= exp(0)$). En effet, intuitivement on a d'autant plus de chances de dire que deux proportions ne sont pas significativement différentes l'une de l'autre que leur rapport (au sens de l'*odds-ratio*) est proche de 1.

Dès lors, il est tout à fait possible de mener une inférence sur la base des *odds-ratio* et de leur intervalle de confiance : si 1 appartient à l'intervalle de confiance à 95 % de l'*odds-ratio*, alors on ne peut pas exclure que la variable explicative et la variable expliquée n'ait aucune relation statistique ; en revanche, si 1 n'appartient pas à l'intervalle de confiance de l'*odds-ratio* à 95 %, alors on peut rejeter l'hypothèse d'absence de relation entre les variables au seuil de 5 %.

Ici 1 appartient à l'intervalle de confiance de `immobi` mais pas de `tresor` ni de `gar_y` (même s'il est tangeant à trois décimales dans le second cas) : on peut donc au seuil de 5 % rejeter l'hypothèse selon laquelle le type d'instrument financier (avance de trésorerie tout particulièrement) et l'existence d'une garantie ne seraient pas liés à la probabilité pour le crédit d'être à taux variable.

b. Utilisez la macro-fonction `%INCLUDE` pour charger le code SAS `logistic_marginal.sas`. Utilisez la macro `%logistic_marginal` pour calculer l'effet marginal moyen associé aux modalités des deux variables qualitatives du modèle `m4` (inspirez-vous de l'exemple présenté dans le support). Comment interprétez-vous ces quantités ?

Proposition de solution

```
/*Code de la macro %logistic_marginal*/
```

```sas
%INCLUDE "\\chemin\vers\le\fichier\logistic_marginal.sas";

/*Utilisation de la macro %logistic_marginal sur le modèle m4*/
%logistic_marginal(
    DATA = m_contran
    , CLASS = categh (REF = 'immobi') gar(REF = 'gar_n') / PARAM = REF
    , MODEL = txvar (DESC) = mt_crdt categh duree_in gar
);
/*NOTES IMPORTANTES : quand on utilise une macro SAS
  - le nom des arguments est suivi de = ;
  - les différents arguments sont séparés par des virgules (et non
     des points-virgules comme les instructions d'une PROC);
  - en général la colorisation syntaxique du logiciel est désactivée.
*/
```

FIGURE 17 – Sortie associée à la question 1.4.b

Variable	Modalite	Effet marginal moyen	Ecart-type	P-valeur
CATEGH	immobi	-0.7817	0.4989	0.1171
CATEGH	tresor	10.9084	1.1464	<.0001
GAR	gar_y	0.0214	0.0043	<.0001

On interprète un effet marginal moyen en régression logistique de façon relativement analogue à l'interprétation d'un coefficient en régression linéaire, sinon que la variable d'intérêt est exprimée sous la forme d'une probabilité.

Ainsi pour la modalité `tresor` de `categh`, l'effet marginal moyen est de 10,9 points de pourcentages : en moyenne dans l'échantillon, les crédits correspondant à des avances de trésorerie ont une probabilité significativement plus élevée de l'ordre de 10,9 points de pourcentages d'être à taux variable par rapport aux crédits associés à des investissements. De même, le fait pour un crédit d'être garanti (plutôt que pas) est associé à une probabilité très légèrement mais néanmoins significativement supérieure d'être à taux variable par rapport aux crédits associés à des investissements. Enfin on n'interprète pas l'effet marginal moyen associé à la modalité `immobi` de `categh` dans la mesure où il n'est pas significatif au seuil de 10 %.

c. Présentez (avec Word par exemple) les résultats du modèle `m4` sous la forme qui vous semble la plus appropriée (libellés des variables, coefficients avec ou sans étoiles, *odds-ratio*, effets marginaux moyens pour les variables qualitatives, etc.).

Cas pratique 1.5 Variables croisées

Dans le dernier cas pratique de cette partie, on aborde la question des variables d'interaction et de leur interprétation. En particulier, on va chercher à répondre à deux questions spécifiques :

— la relation entre le fait que le crédit soit garanti et la probabilité d'être à taux fixe est-elle la même pour toutes les catégories d'instrument financier (trésorerie, investissement et immobilier) ?
— la relation entre durée du crédit et probabilité d'être à taux fixe est-elle la même pour toutes les catégories d'instrument financier ?

a. Construisez manuellement (dans une étape DATA) la variable croisée à intégrer dans un modèle susceptible d'apporter des éléments de réponse à la première des deux questions. Estimez ce modèle.

Proposition de solution

On cherche à estimer un modèle qui autorise le coefficient associé à la variable `gar` à varier avec les modalités de la variable `categh`. En pratique, cela revient à créer la variable qualitative croisée `categh` x `gar` et à l'intégrer dans le modèle.

```
/*Création de la variable croisée categh_gar*/
DATA m_contran;
  SET m_contran;
  categh_gar = COMPRESS(categh||'_'||gar);
RUN;
PROC FREQ DATA = m_contran;
  TABLES categh_gar;
RUN;

/*Estimation du nouveau modèle*/
PROC LOGISTIC DATA = m_contran;
  CLASS categh_gar (REF = 'invest_gar_n') / PARAM = REF;
  MODEL txvar (DESC) = mt_crdt categh_gar duree_in;
RUN;
```

b. Comment interprétez-vous le signe et la significativité des coefficients associés à cette variable croisée ? Changez la modalité de référence pour être en mesure d'interpréter la significativité des écarts entre les coefficients de plusieurs modalités de cette variable.

Proposition de solution

Le signe et la significativité de tous les coefficients de la variable qualitative croisée `categh_gar` doivent être interprétés par rapport à la modalité de référence (ici les crédits liés à un investissement ne disposant pas par ailleurs de garantie).

Ainsi, le coefficient associé à la modalité `invest_gar_y` vaut 1,8553 avec une p-valeur de 0,0009 : **parmi les crédits destinés à des investissements**, le fait d'être garanti est significativement (au seuil de 1 %) associé à une probabilité plus élevée d'être à taux variable, les autres variables du modèle étant égales par ailleurs.

En modifiant la position de la référence, il devient possible d'utiliser le test de significativité pour répondre à la question initialement posée, à savoir déterminer si, au sein d'une catégorie d'instrument financiers donnée, le fait de disposer d'une garantie est associée ou non à la probabilité d'être à taux fixe.

FIGURE 18 – Sortie associée à la question 1.5.b

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-5.2683	0.5385	95.7300	<.0001
mt_crdt		1	0.000277	0.000157	3.1116	0.0777
categh_gar	**immobi_gar_n**	1	0.3448	1.1342	0.0924	0.7611
categh_gar	**immobi_gar_y**	1	0.8554	0.9774	0.7659	0.3815
categh_gar	**invest_gar_y**	1	1.8553	0.5607	10.9511	0.0009
categh_gar	**tresor_gar_n**	1	3.2574	0.5439	35.8728	<.0001
categh_gar	**tresor_gar_y**	1	3.2940	0.5724	33.1163	<.0001
duree_in		1	0.0131	0.00300	19.1962	<.0001

```
/*Changement de la modalité de référence*/
PROC LOGISTIC DATA = m_contran;
  CLASS categh_gar (REF = 'immobi_gar_n') / PARAM = REF;
  MODEL txvar (DESC) = mt_crdt categh_gar duree_in;
RUN;
PROC LOGISTIC DATA = m_contran;
  CLASS categh_gar (REF = 'tresor_gar_n') / PARAM = REF;
  MODEL txvar (DESC) = mt_crdt categh_gar duree_in;
RUN;
```

Ici en l'occurrence :

— **avec `immobi_gar_n` comme référence**, on interprète la significativité du coefficient associé à la modalité `immobi_gar_y` : la p-valeur correspondante est de 0,6937, autrement dit au sein des crédits immobilier, on ne peut pas à un seuil raisonnable écarter l'hypothèse que le fait que le crédit soit garanti ou pas n'entretienne aucune relation avec le fait que le taux soit variable ou non.

— de même **avec `tresor_gar_n` comme référence**, on interprète la significativité du coefficient associé à la modalité `tresor_gar_y` : la p-valeur correspondante est de 0,8807, autrement dit au sein des avances de trésorerie, on ne peut pas à un seuil raisonnable écarter l'hypothèse que le fait que le crédit soit garanti ou pas n'entretienne aucune relation avec le fait que le taux soit variable ou non.

De façon générale, on retient que le fait que le crédit soit garanti semble être associé à une probabilité significativement plus élevée d'être à taux variable uniquement pour les crédits intervenant dans le cadre de projet d'investissement.

c. Ecrivez le modèle qui permette, en introduisant des interactions entre durée du crédit et catégorie d'instrument financier, d'apporter des éléments de réponse à la seconde question. Utilisez l'opérateur `*` dans l'instruction `MODEL` pour que SAS construise

automatiquement les variables croisées correspondantes.

Proposition de solution

Pour répondre à cette seconde question, il convient d'autoriser la valeur du coefficient associé à `duree_in` à varier selon la catégorie d'instrument financier. Pour ce faire, il suffit de croiser ces deux variables de la façon suivante :

$$taux\_variable_i = \beta_0 + \beta_1 montant_i + \beta_2 immobi_i + \beta_3 tresor_i + \beta_4 duree\_in_i + \beta_5 gar_i + \beta_6 immobi_i \times duree\_in_i + \beta_7 tresor_i \times duree\_in_i + \varepsilon_i$$

En pratique dans SAS, il n'est pas impréatif de créer manuellement les variables `immobi` x `duree_in` et `tresor` x `duree_in` : l'opérateur `*` de l'instruction `MODEL` s'en charge automatiquement.

```
/*Introduction d'une interaction categh x duree_in*/
PROC LOGISTIC DATA = m_contran;
  CLASS categh (REF = 'invest') gar(REF = 'gar_n') / PARAM = REF;
  MODEL txvar (DESC) = mt_crdt categh duree_in gar categh*duree_in;
RUN;
```

d. Comment interprétez-vous le signe et la significativité des coefficients associés aux variables correspondant au croisement ? À nouveau modifiez les modalités de référence pour tester les relations entre plusieurs paires de coefficients (pris deux-à-deux).

Proposition de solution

FIGURE 19 – Sortie associée à la question 1.5.d

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-4.5750	0.3995	131.1444	<.0001
mt_crdt		1	0.000262	0.000161	2.6651	0.1026
categh	**immobi**	1	1.5338	0.9369	2.6804	0.1016
categh	**tresor**	1	2.3952	0.4094	34.2260	<.0001
duree_in		1	0.0168	0.00384	19.2050	<.0001
gar	**gar_y**	1	0.4337	0.2026	4.5809	0.0323
duree_in*categh	**immobi**	1	-0.0168	0.00768	4.7749	0.0289
duree_in*categh	**tresor**	1	0.00611	0.00647	0.8916	0.3450

Les coefficients associés aux variables croisées et leur significativité doivent être interprétés par rapport à la modalité de référence de la variable qualitative intervenant dans le croisement, à savoir ici `categh = "invest"`.

Ainsi ici le coefficient associé à la variable `duree_in` (sans croisement) est de 0,168 et celui associé à la variable croisée `duree_in x categh = "immobi"` est égal à -0,168 et est significatif au seuil de 5 % : la relation entre durée initiale du crédit et probabilité d'être à taux fixe est significativement moins forte (au seuil de 5 %) parmi les crédits immobilier comparé aux crédits liés à des investissements, au point qu'elle semble pour ainsi dire totalement décorrélée (0,168 - 0,168 = 0).

À nouveau, changer la modalité de référence permet d'exploiter au mieux les tests de significativité pour répondre aux questions motivant l'analyse.

e. (Complément) Comment dans ce contexte testeriez-vous le caractère statistiquement significatif de la relation entre durée initiale du crédit et probabilité d'être à taux variable **pour les crédits immobiliers** ? Ecrivez le modèle contraint correspondant à ce test et menez-le à bien en utilisant l'instruction `TEST` de la `PROC LOGISTIC`.

Proposition de solution

Dans le modèle

$$taux\_variable_i = \beta_0 + \beta_1 montant_i + \beta_2 immobi_i + \beta_3 tresor_i + \beta_4 duree\_in_i + \beta_5 gar_i + \beta_6 immobi_i \times duree\_in_i + \beta_7 tresor_i \times duree\_in_i + \varepsilon_i$$

la relation entre durée initiale du crédit et probabilité d'être à taux variable est captée par $\beta_4 + \beta_6$. Tester si cette relation est significative revient donc à poser le test statistique :

$$H_0 : \beta_4 + \beta_6 = 0 \quad \text{contre} \quad \beta_4 + \beta_6 \neq 0$$

Ceci est un test d'hypothèse complexe (il fait intervenir plus d'un paramètre) dont le modèle contraint correspondant s'écrit :

$$\begin{aligned} taux\_variable_i &= \beta_0 + \beta_1 montant_i + \beta_2 immobi_i + \beta_3 tresor_i + \beta_4 duree\_in_i + \beta_5 gar_i \\ &\quad - \beta_4 immobi_i \times duree\_in_i + \beta_7 tresor_i \times duree\_in_i + \varepsilon_i \\ &= \beta_0 + \beta_1 montant_i + \beta_2 immobi_i + \beta_3 tresor_i + \beta_4 (duree\_in_i - immobi_i \times duree\_i \\ &\quad + \beta_7 tresor_i \times duree\_in_i + \varepsilon_i \end{aligned}$$

En pratique dans SAS, on estime ce modèle en utilisant l'instruction `TEST` de la `PROC LOGISTIC` :

```
/*Mise en oeuvre d'un test d'hypothèse complexe*/
PROC LOGISTIC DATA = m_contran;
  CLASS categh (REF = 'invest') gar(REF = 'gar_n') / PARAM = REF;
  MODEL txvar (DESC) = mt_crdt categh duree_in gar categh*duree_in;
  TEST duree_in + categhimmobiduree_in = 0;
RUN;
```

La p-valeur du test est de 0,9914 et est donc extrêmement élevée : aux seuils statistiques usuels on ne peut pas rejeter l'hypothèse nulle selon laquelle pour les crédits immobiliers, l'association entre durée initiale du contrat et probabilité d'être à taux variable est nulle.

FIGURE 20 – Sortie associée à la question 1.5.e

Linear Hypotheses Testing Results				
	Label	Wald Chi-Square	DF	Pr > ChiSq
	Test 1	0.0001	1	0.9914

Partie 3 : Adapter la spécification du modèle aux données

Cas pratique 1.6 Modélisation gamma du montant du crédit

On estime un modèle gamma avec fonction de lien logarithmique pour chercher à expliquer le montant d'un crédit en fonction de la catégorie d'instrument financier, de la durée initiale et de la présence ou non d'une garantie.

a. Pourquoi recourir à une régression gamma pour modéliser la variable de montant du crédit plutôt qu'à une régression linéaire classique ? Fournissez des éléments empiriques à l'appui de votre réponse.

Proposition de solution

La variable de montant du crédit est positive, continue et particulièrement asymétrique : sa distribution est extrêmement affectée par quelques valeurs extrêmes qui la rendent particulièrement difficile à modéliser avec un modèle linéaire classique.

```
/*Statistiques descriptives sur la variable mt_crdt*/
PROC UNIVARIATE DATA = m_contran;
    VAR mt_crdt;
RUN;
```

b. Utilisez la `PROC GENMOD` pour estimer le modèle souhaité. Sur le base de l'analyse des effets de Type III, que pensez-vous de la pertinence des variables intégrées dans le modèle ? Interprétez les coefficients en termes d'écart moyen en pourcentages.

Proposition de solution

```
/*Estimation avec la PROC GENMOD*/
PROC GENMOD DATA = m_contran;
    CLASS categh (REF = 'invest') gar(REF = 'gar_n') / PARAM = REF;
    MODEL mt_crdt = categh duree_in gar / DIST = GAMMA LINK = LOG
        TYPE3;
RUN;
```

FIGURE 21 – Sortie associée à la question 1.6.b

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	2.5274	0.1238	2.2849	2.7700	416.99	<.0001
categh	immobi	1	-0.5883	0.2167	-1.0130	-0.1636	7.37	0.0066
categh	tresor	1	1.6490	0.1183	1.4172	1.8808	194.46	<.0001
duree_in		1	0.0241	0.0018	0.0206	0.0276	177.66	<.0001
gar	gar_y	1	0.5800	0.1005	0.3830	0.7771	33.29	<.0001
Scale		1	0.3287	0.0093	0.3110	0.3475		

Note: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
categh	2	200.27	<.0001
duree_in	1	251.76	<.0001
gar	1	35.38	<.0001

Les tests associés à l'analyse des effets de Type III permettent tous de rejeter très largement l'hypothèse d'absence de relation entre les variables introduites dans le modèle et la variable expliquée (p-valeur inférieure à 0,01 dans les trois cas).

En termes d'interprétation des coefficients, on dira par exemple qu'en moyenne dans l'échantillon, les crédits immobiliers sont d'un montant de l'ordre de 59 % inférieur à ceux liés à un investissement, à durée initiale et existence d'une garantie égales par ailleurs. Cette relation est statistiquement significative au seuil de 1 % (p-valeur inférieure à 0,01).

c. Comment interprétez-vous en particulier la valeur du coefficient associé à la modalité `tresor` de `categh` ? Vérifiez par une analyse bivariée que le modèle est bien cohérent avec les données sous-jacentes (et qu'il n'y a pas d'erreur de code !).

Proposition de solution

Le coefficient associé à la modalité `tresor` de `categh` vaut 1,65 et est significatif au seuil de 1 % : en moyenne dans l'échantillon, les avances de trésorerie sont d'un montant de l'ordre de 165 % supérieur à ceux liés à un investissement à durée initiale et existence d'une garantie égale par ailleurs.

Cette valeur particulièrement élevée est confirmée par l'analyse de la distribution jointe de `mt_crdt` et `categh` :

```
/*Distibution jointe de mt_crdt et categh*/
PROC MEANS DATA = m_contran;
```

```
    CLASS categh;
    VAR mt_crdt ;
RUN;

/*Calcul de la moyenne winsorizée à 5 % de mt_crdt selon les
   modalités de categh*/
PROC SORT DATA = m_contran;
  BY categh;
RUN;
PROC UNIVARIATE DATA = m_contran WINSORIZED = 0.05;
    BY categh;
    VAR mt_crdt ;
RUN;
```

L'écart entre moyenne classique et moyenne winsorizée est particulièrement fort pour les avances de trésorerie : les valeurs extrêmes en termes de montant au sein de ce type d'instrument financier risquent d'affecter fortement la valeur des estimateurs dans le modèle de régression.

d. Réestimez le modèle en laissant de côté les 5 valeurs les plus élevées de `mt_crdt`, puis les 5 % de valeurs les plus élevées de `mt_crdt`. Cela modifie-t-il substantiellement les résultats de la modélisation ?

Proposition de solution

```
/*Création de variables indicatrices de filtre*/
DATA m_contran;
  SET m_contran;
  drop5 = (mt_crdt >= 4300);
  drop5pct = (mt_crdt >= 272);
RUN;

/*Estimation en excluant les 5 observations les plus extrêmes*/
PROC GENMOD DATA = m_contran;
  WHERE drop5 = 0;
  CLASS categh (REF = 'invest') gar(REF = 'gar_n') / PARAM = REF;
  MODEL mt_crdt = categh duree_in gar / DIST = GAMMA LINK = LOG TYPE3;
RUN;

/*Estimation en excluant les 5 pourcents d'observations les plus
   extrêmes*/
PROC GENMOD DATA = m_contran;
  WHERE drop5pct = 0;
  CLASS categh (REF = 'invest') gar(REF = 'gar_n') / PARAM = REF;
  MODEL mt_crdt = categh duree_in gar / DIST = GAMMA LINK = LOG TYPE3;
RUN;
```

En excluant les 5 observations les plus extrêmes, l'ampleur du coefficient associé à la modalité `tresor` de `categh` diminue (à 0,6442) ; en excluant les 5 pourcents des observations les plus extrêmes, le signe du coefficient associé à la modalité `tresor` de `categh` change (avec -0,1351) et n'est plus significativement différent de 0 au seuil de 5 %.

Cas pratiques à partir de l'enquête PISA 2012

Nota bene *Les instructions des cas pratiques de cette partie sont volontairement moins directives : ils doivent vous permettre de davantage expérimenter sur des données différentes de celles de la table M_CONTRAN. Selon le rythme de progression de chacun, ces cas pratiques ont vocation à être traités pendant ou à l'issue de la formation.*

L'enquête Pisa (*Program for International Student Assessment*) est une enquête réalisée **tous les trois ans** par l'Organisation de coopération et de développement économique (OCDE) dans une soixantaine de pays auprès des **élèves de 15 ans** (quelle que soit leur classe au moment de l'enquête).

Elle vise à mesurer les **acquis des élèves de 15 ans dans trois disciplines** : mathématiques, compréhension de l'écrit (ou *littératie*) et sciences. En plus des scores aux **tests standardisés** de mathématiques, compréhension de l'écrit et sciences, cette enquête comporte de très nombreuses informations sur l'origine sociale des élèves, leurs conditions d'enseignement ainsi que leur rapport aux enseignants et à l'école.

Les données nécessaires sont disponibles ici et les questionnaires adressés aux élèves et aux établissements ici. L'ensemble a été librement téléchargé à partir de cette page.

Variable	**Description**
cnt	Pays
stidstd	Identifiant de l'élève
schoolid	Identifiant de l'établissement
w_fstuwt	Poids de sondage final de l'élève
st01q01	Classe en nombre d'années depuis l'entrée en primaire : la $10^{ème}$ classe correspond à la seconde en France.
st04q01	Sexe : (1) Femme (2) Homme
st05q01	A suivi une scolarité pré-primaire (1) Non (2) Oui, un an ou moins (3) Oui, plus d'un an
st07q01 st07q02 st07q03	A redoublé à un moment de sa scolarité : (1) Non (2-3) Oui, une ou plusieurs fois
st08q01	Est arrivé en retard au cours des deux semaines précédant l'enquête
st09q01	A séché les cours au cours des deux semaines précédant l'enquête
anxmat	Score synthétique d'anxiété en mathématiques
disclima	Score synthétique de climat de discipline dans la classe
escs	Indicateur synthétique de statut économique, social et culturel

Variable	Description
immig	Immigration : (1) Né en France de parents nés en France (2) Immigré de deuxième génération (3) Immigré de première génération
hisced	Niveau d'étude le plus élevé des parents (nomenclature CITE)
pv1math	Score synthétique à l'évaluation de mathématiques
pv1read	Score synthétique à l'évaluation de compréhension de l'écrit
pv1scie	Score synthétique à l'évaluation de sciences
sc01q01	Statut public ou privé de l'établissement (1) public (2) privé
sc03q01	Taille de la commune de l'établissement : (1) *Village* (2) *Small town* (3) *Town* (4) *City* (5) *Large city*
sc05q01	Taille de la classe en cours de français : (01) 15 ou moins (02) 16-20 (03) 21-25 ... (08) 46-50 (09) Plus de 50 élèves

Cas pratique 2.1 Analyse multivariée du retard scolaire

Cette première proposition de cas pratique vise à rechercher quelques déterminants du retard scolaire (au sens du fait d'avoir redoublé à un moment ou à un autre au cours de scolarité), parmi lesquels notamment : le sexe, le fait d'avoir suivi une scolarité pré-primaire, le statut économique, social et culturel, le niveau d'études le plus élevé des parents, le statut public ou privé de l'établissement, la taille de la classe en cours de français, etc.

a. Les variables `st07q01`, `st07q02` et `st07q03` renseignent sur le fait d'avoir redoublé (1 : Non, 2 : Une fois, 3 : Plus d'une fois) à plusieurs moments de la scolarité :
   - `st07q01` : au cours de la primaire (ISCED 1) ;
   - `st07q02` : au cours du collège (ISCED 2) ;
   - `st07q03` : au cours du lycée (ISCED 3).

   Construisez une variable synthétisant le fait qu'un élève ait, à un moment ou à un autre de sa scolarité, redoublé une ou plusieurs fois.

   Proposition de solution

```
/*Construction de la variable retard*/
DATA pisa12;
    SET pisa12;
    retard = ( st07q01 IN(2, 3) OR st07q02 IN(2, 3) OR st07q03 IN(2,
       3) );
RUN;
```

```
PROC FREQ DATA = pisa12;
    TABLES retard;
RUN;
/*Environ 27 % de l'échantillon est en retard scolaire*/
```

b. (i) Analysez la distribution des variables de sexe, de scolarisation pré-primaire et de statut social, économique et culturel.

Proposition de solution

```
/*Analyse univariée de la variable de sexe*/
PROC FREQ DATA = pisa12;
        TABLES st04q01;
RUN;
/*Note : les filles sont codées 1 et les garçons 2*/

/*Remarque : en pondérant les résultats sont très proches*/
PROC FREQ DATA = pisa12;
        TABLES st04q01;
        WEIGHT w_fstuwt;
RUN;

/*Analyse univariée de la variable de scolarisation pré-primaire*/
PROC FREQ DATA = pisa12;
    TABLES st05q01;
RUN;
/*Les trois modalités sont très déséquilibrées.
On décide de recoder ensemble les enfants qui n'ont
pas connu une scolarisation pré-primaire complète.*/
DATA pisa12;
    SET pisa12;
    preprim = (st05q01 = '3');
RUN;
PROC FREQ DATA = pisa12;
    TABLES preprim;
RUN;

/*Analyse univariée de la variable de statut économique, social et
   culturel*/
PROC FREQ DATA = pisa12;
    TABLES escs;
RUN;
/*Il semble que cette variable soit en fait de nature quantitative :
c'est un score synthétisant le statut économique, social et culturel.
On l'analyse donc avec la PROC UNIVARIATE.*/
PROC UNIVARIATE DATA = pisa12;
    VAR escs;
RUN;
/*Ce score est manifestement calibré de façon que ses quantiles
correspondent à certaines valeurs particulières (Q1 = -0,60, Q3 =
   0,60).*/

/*Pour le rendre plus explicite et interprétable,
on le dichotomise en quintiles.*/
```

```
PROC RANK DATA = pisa12 OUT = pisa12 GROUPS = 5;
    VAR escs;
    RANKS escsq;
RUN;
/*On vérifie que la PROC RANK a bien fait ce que l'on souhaitait*/
PROC MEANS DATA = pisa12;
    CLASS escsq;
    VAR escs;
RUN;
```

(ii) Utilisez ensuite les outils de la statistique bivariée pour mettre en évidence une éventuelle relation entre chacune de ces variables prise indépendamment et le retard scolaire.

Proposition de solution

```
/*Analyse de la relation entre sexe et retard scolaire*/
PROC FREQ DATA = pisa12;
  TABLES st04q01 * retard / CHISQ CELLCHI2 NOCOL;
RUN;
```

30,21 % des garçons ont redoublé au moins une fois contre 27,25 % des élèves dans l'ensemble de l'échantillon. Ce résultat semble ainsi indiquer une certaine sur-représentation des élèves ayant redoublé parmi les garçons.

La p-valeur du test d'indépendance du χ^2 est inférieure à 0,01 aussi on peut rejeter au seuil de 1 % l'hypothèse que ces deux variables ne sont pas liées.

Il ressort donc de cette première analyse que sexe et retard scolaire sont statistiquement liés : les garçons sont sensiblement plus souvent en situation de retard scolaire que les filles.

```
/*Analyse de la relation entre scolarisation pré-primaire et retard
   scolaire*/
PROC FREQ DATA = pisa12;
  TABLES preprim * retard / CHISQ CELLCHI2 NOCOL;
RUN;
```

42,98 % des élèves n'ayant pas suivi une scolarité pré-primaire complète ont redoublé au moins une fois contre 27,25 % des élèves dans l'ensemble de l'échantillon. Ce résultat semble ainsi indiquer une nette sur-représentation des élèves ayant redoublé parmi les élèves n'ayant ps suivi une scolarité pré-primaire complète.

La p-valeur du test d'indépendance du χ^2 est inférieure à 0,01 aussi on peut rejeter au seuil de 1 % l'hypothèse que ces deux variables ne sont pas liées.

Il ressort donc de cette première analyse que scolarité pré-primaire et retard scolaire sont statistiquement liés : les élèves n'ayant pas suivi une scolarité pré-primaire complète sont sensiblement plus souvent en situation de retard scolaire que les autres.

```
/*Analyse de la relation entre statut économique, social et culturel
   et retard scolaire*/
PROC FREQ DATA = pisa12;
  TABLES escsq * retard / CHISQ CELLCHI2 NOCOL;
RUN;
```

8,77 % des élèves au statut économique et culturel le plus élevé ont redoublé au moins une fois contre 27,16 % des élèves dans l'ensemble de l'échantillon (pour lesquels la variable de statut économique, social et culturel est renseignée). Inversement, 48,72 % des élèves au statut économique, social et culturel le plus faible ont redoublé au moins une fois. Ces éléments traduisent une nette sur-représentation des élèves ayant redoublé parmi les élèves au statut économique, social et culturel le plus faible.

La p-valeur du test d'indépendance du χ^2 est inférieure à 0,01 aussi on peut rejeter au seuil de 1 % l'hypothèse que ces deux variables ne sont pas liées.

Il ressort donc de cette première analyse que statut économique, social et culturel et retard scolaire sont statistiquement liés : les élèves dont le statut économique, social et culturel est le plus faible sont sensiblement plus souvent en situation de retard scolaire que les autres.

(iii) Intégrez enfin l'ensemble de ces variables dans un modèle de régression logistique multiple : jugez de la qualité de la modélisation, interprétez le signe et la significativité des coefficients et réexprimez les coefficients des variables qualitatives en termes d'*odds-ratio* ou d'effets marginaux moyens.

Proposition de solution

```
/*Modèle de régression logistique multiple*/
ODS GRAPHICS ON;
PROC LOGISTIC DATA = pisa12 PLOT(ONLY) = ROC;
  CLASS st04q01(REF = '2') preprim(REF = '1') escsq(REF = '2') /
     PARAM = REF;
  MODEL retard (DESC) = st04q01 preprim escsq;
RUN;
```

Etant donné son extrême parcimonie (6 variables indicatrices), ce modèle est plutôt bon : il atteint un pourcentage de concordance de 68,3 % et une aire sous la courbe ROC de 0,7190. Surtout, l'ensemble des tests de significativité conduisent à des p-valeurs inférieures à 0,01 % (c'est donc *a fortiori* le cas du test de significativité globale par le ratio de vraisemblance) : ce modèle rassemble des facteurs extrêmement explicatifs du retard scolaire.

Le fait de ne pas avoir suivi de scolarité pré-primaire complète et le statut économique, social et culturel apparaissent comme les variables dont l'association avec le retard scolaire est la plus forte. Pour le statut économique, social et culturel en particulier, chacun des quintiles recodés précédemment semble associé à une probabilité très spécifique de retard scolaire, ce que l'on avait déjà observé dans les statistiques bivariées. Cette analyse confirme également qu'à statut économique, social et culturel et scolarisation pré-primaire égaux par ailleurs, les garçons ont une probabilité siginifcativement plus élevée (au seuil de 1 %) d'être en retard scolaire que les filles, même si l'ampleur de l'effet est moindre que pour les autres variables de la modélisation.

Plus concrètment, à sexe et scolarisation pré-primaire égaux par ailleurs, un élève au statut économique, social et culturel le plus faible à 2,7 fois plus de chances d'être en retard scolaire qu'un élève de statut intermédiaire, et un élève de statut le plus élevé a 1 / 0,285 = 3,5 fois moins de chances d'être en retard scolaire qu'un élève de statut intermédiaire.

c. On s'intéresse tout particiculièrement à la variable `immig` codant l'origine migratoire des élèves (et qui distingue les individus immigrés de première ou de deuxième génération).

(i) Recodez cette variable en deux modalités ("Né en France de parents nés en France" *versus* "Immigré ou enfant d'immigré") et intégrez-la comme seul variable explicative du retard scolaire dans une régression logistique simple.

Proposition de solution

```
/*Travail sur la variable d'origine géographique*/
PROC FREQ DATA = pisa12;
  TABLES immig;
RUN;
DATA pisa12;
  SET pisa12;
  immig2 = ( immig IN(2, 3) );
RUN;

/*Régression logistique simple*/
PROC LOGISTIC DATA = pisa12;
    CLASS immig2(REF = '0') / PARAM = REF;
    MODEL retard (DESC) = immig2;
RUN;
```

Le coefficient associé à la variable `immig2` dans cette régression est positif et significativement différent de 0 au seuil de 1 % : le modèle indique qu'il existe une relation significative et positive entre origine migratoire et retard scolaire.

On peut tout de suite noter que ce modèle est particulièrement frustre : une seule variable explicative (pas plus explicatif qu'un tri croisé donc, aucun contrôle des effets de structure), seulement 19 % de concordance.

(ii) Intégrez ensuite le statut social, économique et culturel ainsi que le fait d'avoir suivi une scolarité pré-primaire. Que constatez-vous quand à l'évolution du coefficient associé au fait d'être immigré ou enfant d'immigré entre les deux modèles ? Comment comprenez-vous ce phénomène ?

Proposition de solution

```
/*Régression logistique multiple*/
PROC LOGISTIC DATA = pisa12;
    CLASS immig2(REF = '0') escsq(REF = '2') preprim(REF = '1')/
       PARAM = REF;
    MODEL retard (DESC) = immig2 escsq preprim;
RUN;
```

Les deux variables ajoutées augmentent considérablement le pouvoir prédictif du modèle : il est donc indispensable de les intégrer au modèle. Le principal résultat de ce nouveau modèle est que l'amplitude du coefficient associé à la variable immig2 est beaucoup plus faible une fois contrôlé par les effets de structure captés par le statut

économique, social et culcturel et la scolarisation pré-primaire. Ce coefficient reste cependant significatif au seuil de 1 %.

Ce résultat doit nous inviter à rechercher dans les données de l'enquête si d'éventuels autres effets de structure ne viendraient pas biaiser la valeur du coefficient associé à l'origine migratoire, pour affiner encore l'analyse et mesurer la force effective de son association avec retard scolaire.

Cas pratique 2.2 Analyse multivariée de l'anxiété en mathématiques

La variable `anxmat` est un indicateur synthétique de l'anxiété vis-à-vis des mathématiques. On cherche ici à identifier certains déterminants de l'anxiété en mathématiques chez les élèves **particulièrement anxieux**.

a. Menez l'analyse univariée de la variable `anxmat`. Que constatez-vous quant aux non-réponses ? Recodez cette variable de façon à isoler les élèves particulièrement anxieux en mathématiques (score d'anxiété supérieur au troisième quartile de la distribution).

Proposition de solution

```
/*Analyse univariée de la variable anxmat*/
PROC UNIVARIATE DATA = pisa12;
    VAR anxmat;
RUN;
```

Les non-réponses sont nombreuses (1/3 de l'échantillon). Pour pouvoir exploiter cette variable, il faut faire l'hypothèse que les individus n'ayant pas répondu n'ont pas de caractéristiques spécifiques eu égard à l'anxiété en mathématiques que les autres (ce qui n'est pas nécessairement évident).

```
/*Dichotomisation de la variable en utilisant
le troisième quartile comme seuil*/
DATA pisa12;
    SET pisa12;
    IF anxmat NE . THEN anxmat2 = (anxmat  > 0.79);
RUN;
PROC FREQ DATA = pisa12;
    TABLES anxmat2;
RUN;
```

b. Menez une régression logistique multivariée sur l'anxiété en mathématiques, en cherchant à répondre à plusieurs questions :
   — Garçons et filles sont-ils également anxieux en mathématiques ?
   — Les élèves en retard scolaire sont-ils davantage anxieux que les autres ?
   — Établissements publics et privés diffèrent-ils sensiblement dans le niveau d'anxiété de leurs élèves en mathématiques ?
   — Comment les performances en mathématiques sont-elles liées à l'anxiété vis-à-vis de cette matière ?

— Sexe et performances en mathématiques interagissent-ils pour expliquer le niveau d'anxiété dans cette matière ?

Proposition de solution

La plupart des variables auxquelles il est fait allusion ont déjà été utilisées au cas pratique précédent, si ce n'est la variable de performance en mathématiques `pv1math`. Cette variable est un score synthétique dont les caractéristiques sont calibrées (par exemple la moyenne de tous les pays de l'OCDE aut 500). Afin de pouvoir l'interpréter plus facilement et comme précédemment avec le statut économique, social et culturel, on recode cette variable en quintiles :

```
PROC RANK DATA = pisa12 GROUPS = 5 OUT = pisa12;
    VAR pv1math;
    RANKS pv1mathq;
RUN;
PROC MEANS DATA = pisa12;
    CLASS pv1mathq;
    VAR pv1math;
RUN;
```

On peut alors mener une première régression logistique, sans variable d'interaction :

```
PROC LOGISTIC DATA = pisa12;
    CLASS st04q01(REF = '2') retard(REF = '0') sc01q01(REF = '1')
       pv1mathq(REF = '2') / PARAM = REF;
    MODEL anxmat2 (DESC) = st04q01 retard sc01q01 pv1mathq;
RUN;
```

Le test du ratio de vraisemblance de significativité globale indique que ce modèle est dans l'ensemble explicatif (p-valeur du test inférieure à 0,01), et par ailleurs son pourcentage de concordance est satisfaisant avec 65,7 % (même si on souhaiterait dépasser les 70 % pour bien faire). Assez logiquement, c'est la variable de résultat en mathématiques qui présente la plus forte association avec l'anxiété ressentie vis-à-vis de cette matière, les meilleurs élèves ressentant significativement moins d'anxiété (probabilité 3,3 fois plus faible d'être anxieux en mathématiques pour les 20 % des élèves les meilleurs par rapport au élèves de niveau moyen).

Le principal enseignement de ce modèle est que le sexe est également extrêmement lié à l'anxiété en mathématique, les filles étant, à résultat, statut de l'établissement et retard scolaire égaux par ailleurs significativement (au seuil de 1 %) davantage susceptible de ressentir de l'anxiété vis-à-vis des mathématiques (probabilité de l'ordre de 2 fois supérieure par rapport aux garçons).

Plus surprenant, le statut de l'établissement et le retard scolaire semblent également liés à l'anxiété en mathématiques :

— au seuil de 5 %, les élèves scolarisés dans un établissement privé se disent plus souvent anxieux en mathématiques (à sexe, résultat et retard scolaire égaux par ailleurs) que les élèves scolarisés dans un établissement public ;

— au seuil de 10 %, les élèves en retard scolaire se disent moins souvent anxieux en mathématiques (à sexe, résultat et statut de l'établissement égaux par ailleurs) que les élèves "à l'heure" scolairement. Ce résultat est néanmoins fragile et demande à être confirmé.

Dans un second modèle, on cherche à approfondir la relation entre sexe et anxiété en mathématiques en croisant sexe et résultat en mathématiques :

```
PROC LOGISTIC DATA = pisa12;
    CLASS st04q01(REF = '2') retard(REF = '0') sc01q01(REF = '1')
       pv1mathq(REF = '2') / PARAM = REF;
    MODEL anxmat2 (DESC) = st04q01 retard sc01q01 pv1mathq st04q01 *
       pv1mathq;
RUN;
```

Ce modèle complémentaire n'est pas très concluant : il n'est pas beaucoup plus explicatif ni prédictif que le précédent (pourcentage de concordance à peine supérieur) et conduit à des p-valeurs pour les tests de significativité des coefficients relativement élevées.

Le seul élément véritablement intéressant tient au fait que, pour certaines plages de résultats en mathématiques (légèrement inférieurs à la moyenne ou très supérieurs à la moyenne), les filles présentent un niveau d'anxiété sensiblement plus élevé que les garçons (au seuil de 10 %). Ces résultats demandent néanmoins à être confirmés par des analyses complémentaires.