La régression sur échantillon avec SAS

Josiane LE GUENNEC

INSEE, CEPE-ENSAI

Introduction

Régression linéaire, analyse de la variance et régression logistique se pratiquent, avec SAS, au moyen des procédures REG, GLM, LOGISTIC. Appropriées à la recherche de relations linéaires dans une population, ces procédures ont l'inconvénient de ne pas estimer la précision de façon correcte lorsque les données analysées proviennent d'un échantillon aléatoire. La possibilité de pondérer les observations, notamment par leur poids de sondage, permet bien d'obtenir le bon estimateur d'un coefficient de régression, mais le calcul de variance ne tient aucun compte du plan de sondage, et estime donc imparfaitement le caractère significatif ou non d'une variable du modèle.

Les nouvelles versions du logiciel ont comblé cette lacune. La procédure SURVEYREG, introduite dès la version 8, réalise la régression linéaire, l'analyse de la variance et de la covariance sur des données d'échantillon. La procédure SURVEYLOGISTIC, apparue dans la version 9, effectue la régression logistique sur échantillon aléatoire. Dans ces deux nouvelles procédures, la variance est estimée en tenant compte du plan de sondage, ce qui permet une appréciation plus précise de l'apport d'une variable exogène dans un modèle linéaire.

On présentera successivement :

- la régression linéaire
- l'analyse de la variance
- la régression logistique

en montrant dans chaque cas :

- les méthodes de calcul mises en œuvre par le logiciel
- les différences que l'on peut attendre de l'usage de SURVEYREG ou de SURVEYLOGISTIC, par rapport aux procédures REG, GLM ou LOGISTIC appliquées au même échantillon, à partir des données provenant d'une enquête par sondage.

1. La régression linéaire

On a tiré un échantillon aléatoire s de taille n dans la population U, selon un plan de sondage quelconque. Chaque unité sélectionnée a un poids de sondage w_k égal à l'inverse de sa probabilité d'inclusion π_k. Le questionnaire d'enquête relève, pour chaque individu k de l'échantillon, la valeur des variables quantitatives Y et X_1 à X_p.

1.1. Rappels théoriques

On suppose que, dans la population, la variable Y est liée aux variables X_1 à X_p par une relation linéaire. Pour un individu *k*, celle-ci s'écrit :

$$Y_k = \sum_{j=1}^{p} b_j X_{jk} + U_k = \mathsf{b}'\mathsf{X}_k + U_k \qquad (1)$$

On souhaite l'estimer à partir des données de l'échantillon, la valeur de ces variables n'étant pas connue sur le reste de la population.

Soit :

$$\tilde{\mathsf{b}}' = \left[\tilde{b}_1 ... \tilde{b}_j ... \tilde{b}_p\right] = (\mathsf{X}'\mathsf{X})^{-1}\mathsf{X}'\mathsf{Y} \qquad (2)$$

le vecteur des coefficients de régression estimés dans la population entière. Le problème se ramène à estimer, dans l'échantillon, la valeur des coefficients $\tilde{b}_j$.

On note :

- $\mathsf{Y}_s' = [y_1 ... y_k ... y_n]$, $\mathsf{X}_{ks}' = [x_{1k} ... x_{jk} ... x_{pk}]$ et $\mathsf{X}_s(n,p)$ les valeurs observées sur l'échantillon.
- $\tilde{\mathsf{U}}_s' = [\tilde{u}_1 ... \tilde{u}_k ... \tilde{u}_n] = \left[\left(y_1 - \mathsf{X}'_{1s}\tilde{\mathsf{b}}\right)..\left(y_k - \mathsf{X}'_{ks}\tilde{\mathsf{b}}\right)..\left(y_n - \mathsf{X}'_{ns}\tilde{\mathsf{b}}\right)\right] = \mathsf{Y}_s' - \tilde{\mathsf{b}}'\mathsf{X}_s'$ est le vecteur des résidus de la régression estimée dans la population, pour les unités de l'échantillon.
- $\hat{\mathsf{b}}' = \left[\hat{b}_1 ... \hat{b}_j ... \hat{b}_p\right]$ est le vecteur des coefficients de régression estimés par MCO dans l'échantillon.
- $\hat{\mathsf{U}}' = [\hat{u}_1 ... \hat{u}_k ... \hat{u}_n] = \left[\left(y_1 - \mathsf{X}'_{1s}\hat{\mathsf{b}}\right)... \left(y_k - \mathsf{X}'_{ks}\hat{\mathsf{b}}\right)... \left(y_n - \mathsf{X}'_{ns}\hat{\mathsf{b}}\right)\right] = \mathsf{Y}_s' - \hat{\mathsf{b}}'\mathsf{X}_s'$ est le vecteur des résidus de la régression estimée dans l'échantillon.
- $\mathsf{W} = Diag(w_k) = Diag\left(\frac{1}{\pi_k}\right)$ est la matrice diagonale des poids de sondage des unités de l'échantillon.

1.1.1. L'estimateur $\hat{\mathsf{b}}$ des coefficients de régression

L'expression (2) des coefficients de régression estimés dans l'ensemble de la population peut s'écrire comme le produit de l'inverse d'une matrice T (*p*,*p*) avec un vecteur t (*p*,1) :

$$\tilde{\mathsf{b}} = \mathsf{T}^{-1}\mathsf{t} \qquad (3)$$

avec : $\mathsf{T} = \mathsf{X}'\mathsf{X}$ $\qquad$ $\mathsf{t} = \mathsf{X}'\mathsf{Y}$

Chaque élément α_{ij} de la matrice T a pour expression : $\alpha_{ij}(\mathsf{T}) = \sum_{k\in U} X_{ik}X_{jk}$.

Chaque élément β_j du vecteur t a pour expression : $\beta_j(\mathsf{t}) = \sum_{k\in U} X_{jk}Y_k$.

Dans l'échantillon s, les valeurs α_{ij} et β_j peuvent être estimées par la méthode d'Horvitz-Thomson :

$$\hat{\alpha}_{ij} = \sum_{k\in \mathsf{s}} \frac{x_{ik}x_{jk}}{\pi_k} \qquad \hat{\beta}_j = \sum_{k\in \mathsf{s}} \frac{x_{jk}y_k}{\pi_k}$$

On en déduit les estimateurs sans biais de la matrice T et du vecteur t :

$$\hat{\mathsf{T}} = \left[\hat{\alpha}_{ij}\right] = \mathsf{X_s'WX_s}$$
$$\hat{\mathsf{t}} = \left[\hat{\beta}_{j}\right] = \mathsf{X_s'WY_s}$$

Les *p* coefficients de régression $\hat{b}_j$ sont par conséquent estimés dans l'échantillon par :

$$\boxed{\hat{\mathsf{b}} = \hat{\mathsf{T}}^{-1}\hat{\mathsf{t}} = \left(\mathsf{X_s'WX_s}\right)^{-1}\mathsf{X_s'WY_s}} \qquad (4)$$

1.1.2. Espérance de $\hat{\mathsf{b}}$

Si $\hat{\mathsf{T}}$ estime sans biais la matrice T, $\hat{\mathsf{T}}^{-1}$ est un estimateur biaisé de la matrice inverse T^{-1}. On obtient une expression approchée de l'espérance et de l'erreur quadratique moyenne de l'estimateur $\hat{\mathsf{b}}$ en linéarisant le premier membre de (4) par la méthode de Taylor à l'ordre 1 :

$$\hat{\mathsf{b}} - \tilde{\mathsf{b}} \approx \frac{\delta\tilde{\mathsf{b}}}{\delta\mathsf{t}}\left(\hat{\mathsf{t}} - \mathsf{t}\right) + \frac{\delta\tilde{\mathsf{b}}}{\delta\mathsf{T}}\left(\hat{\mathsf{T}} - \mathsf{T}\right) = \mathsf{T}^{-1}\left(\hat{\mathsf{t}} - \mathsf{t}\right) - \mathsf{T}^{-1}\left(\hat{\mathsf{T}} - \mathsf{T}\right)\mathsf{T}^{-1}\mathsf{t} = \mathsf{T}^{-1}\left(\hat{\mathsf{t}} - \hat{\mathsf{T}}\tilde{\mathsf{b}}\right) \qquad (5)$$

Il résulte de (5), en utilisant la relation (3) : $\boxed{E\left(\hat{\mathsf{b}}\right) \approx \tilde{\mathsf{b}}}$

$\hat{\mathsf{b}}$ est un estimateur approximativement sans biais des coefficients de régression $\tilde{\mathsf{b}}$ calculés dans la population.

1.1.3. Erreur quadratique moyenne de $\hat{\mathsf{b}}$

$$EQM\left(\hat{\mathsf{b}}\right) = E\left(\hat{\mathsf{b}} - \tilde{\mathsf{b}}\right)\left(\hat{\mathsf{b}} - \tilde{\mathsf{b}}\right)' \approx V\left[\mathsf{T}^{-1}\left(\hat{\mathsf{t}} - \hat{\mathsf{T}}\tilde{\mathsf{b}}\right)\right] = \mathsf{T}^{-1}\left[V\left(\hat{\mathsf{t}} - \hat{\mathsf{T}}\tilde{\mathsf{b}}\right)\right]\mathsf{T}^{-1}$$
$$\hat{\mathsf{t}} - \hat{\mathsf{T}}\tilde{\mathsf{b}} = \mathsf{X_s'WY_s} - \mathsf{X_s'WX_s}\tilde{\mathsf{b}} = \mathsf{X_s'W}\left(\mathsf{Y_s} - \mathsf{X_s}\tilde{\mathsf{b}}\right) = \mathsf{X_s'W}\tilde{\mathsf{U}}_\mathsf{s}$$

où $\tilde{\mathsf{U}}_\mathsf{s}$ est le vecteur $(n,1)$ réduit à l'échantillon s des résidus $\tilde{u}_k$ de la régression de Y sur X dans la population.

Le vecteur $\left(\hat{\mathsf{t}} - \hat{\mathsf{T}}\tilde{\mathsf{b}}\right)$ a pour éléments : $\hat{Z}_j = \sum_{k\in\mathsf{s}} \frac{x_{jk}\tilde{u}_k}{\pi_k}$, estimateur Horvitz-Thomson du total Z_j dans la population de la variable définie par : $Z_{jk} = X_{jk}\tilde{U}_k$. Sa variance est le vecteur dont les éléments sont les variances des estimateurs de totaux $\hat{Z}_{jk}$.

Les coefficients de régression estimés dans l'échantillon ont donc pour variances et covariances approchées :

$$\boxed{EQM\left(\hat{b}\right) \approx \left(\mathsf{X'X}\right)^{-1}\left[\mathsf{V}\right]\left(\mathsf{X'X}\right)^{-1}}$$

avec :

$$V_j = V\left(\hat{Z}_j\right) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{Z_{jk}}{\pi_k}\frac{Z_{jl}}{\pi_l} = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{X_{jk}\tilde{U}_k}{\pi_k}\frac{X_{jl}\tilde{U}_l}{\pi_l}$$

$$V_{ij} = Cov\left(\hat{Z}_i, \hat{Z}_j\right) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{Z_{ik}}{\pi_k}\frac{Z_{jl}}{\pi_l} = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{X_{ik}\tilde{U}_k}{\pi_k}\frac{X_{jl}\tilde{U}_l}{\pi_l}$$

Chaque élément de la matrice V ci-dessus est estimé par l'estimateur Horvitz-Thomson de la variance d'un total. L'erreur quadratique moyenne des coefficients de régression estimés dans l'échantillon est donc elle-même estimée par :

$$\boxed{E\hat{Q}M\left(\hat{\mathsf{b}}\right)=\hat{\mathsf{T}}^{-1}\left[\hat{\mathsf{V}}\left(\hat{\mathsf{t}}-\hat{\mathsf{T}}\tilde{\mathsf{b}}\right)\right]\hat{\mathsf{T}}^{-1}=\left(\mathsf{X}_s'\mathsf{W}\mathsf{X}_s\right)^{-1}\hat{\mathsf{V}}\left(\mathsf{X}_s'\mathsf{W}\mathsf{X}_s\right)^{-1}} \qquad (6)$$

avec :

$$\hat{\mathsf{V}}=\left[\hat{V}_j \ldots \hat{V}_{ij}\right]$$

$$\hat{V}_j=\hat{V}\left(\hat{Z}_j\right)=\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{z_{jk}}{\pi_k}\frac{z_{jl}}{\pi_l}=\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{x_{jk}\hat{u}_k}{\pi_k}\frac{x_{jl}\hat{u}_l}{\pi_l}$$

$$\hat{V}_{ij}=C\hat{o}v\left(\hat{Z}_i,\hat{Z}_j\right)=\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{z_{ik}}{\pi_k}\frac{z_{jl}}{\pi_l}=\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{x_{ik}\hat{u}_k}{\pi_k}\frac{x_{jl}\hat{u}_l}{\pi_l} \qquad (7)$$

où $\hat{u}_k$ est le résidu de la régression dans l'échantillon de Y_s sur X_s : $\hat{u}_k=y_k-\hat{y}_k=y_k-\sum_{j=1}^{p}\hat{b}_j x_{jk}$.

1.1.4. Equation d'analyse de la variance et coefficient de détermination

La somme des écarts à la moyenne de la variable d'intérêt est mesurée dans l'échantillon et se décompose ainsi :

$$\sum_{k\in s}\frac{\left(y_k-\hat{\bar{Y}}\right)^2}{\pi_k}=\sum_{k\in s}\frac{\left(\hat{y}_k-\hat{\bar{Y}}\right)^2}{\pi_k}+\sum_{k\in s}\frac{\left(\hat{u}_k\right)^2}{\pi_k}$$

$$\hat{R}^2=\frac{\sum_{k\in s}\frac{\left(\hat{y}_k-\hat{\bar{Y}}\right)^2}{\pi_k}}{\sum_{k\in s}\frac{\left(y_k-\hat{\bar{Y}}\right)^2}{\pi_k}}=1-\frac{\sum_{k\in s}\frac{\left(\hat{u}_k\right)^2}{\pi_k}}{\sum_{k\in s}\frac{\left(y_k-\hat{\bar{Y}}\right)^2}{\pi_k}}$$

1.1.5. Test de nullité d'un paramètre

Pour tester l'hypothèse nulle Ho : $b_j=0$, on substitue, au numérateur de la fonction pivotante habituelle, l'estimateur $\hat{b}_j$ et au dénominateur la racine carrée de sa variance estimée en (6) et (7) compte tenu du plan de sondage: $\hat{T}=\frac{\hat{b}_j}{\hat{\sigma}_{\hat{b}_j}}$. On admet alors que ce rapport suit approximativement, sous Ho, une loi de Student à $(n-1)$ degrés de liberté lorsque le sondage comprend un seul degré et n'est pas stratifié, à $(n-H)$ degrés de liberté dans le cas d'un sondage stratifié à un degré, où *H* est le nombre de strates.

1.2. La procédure SURVEYREG du logiciel SAS

1.2.1. Le coefficient de régression estimé

La procédure SURVEYREG estime les coefficients de régression entre une variable numérique Y (variable dépendante) et p variables explicatives quantitatives $X_1 \dots X_j \dots X_p$, par la méthode des moindres carrés ordinaires, selon la formule (4) :

$$\boxed{\hat{\beta}_{SAS} = \left(X_s' W X_s\right)^{-1} X_s' W Y_s} \qquad \text{où} \quad W = Diag(w_k)$$

Les observations de l'échantillon sont pondérées par leurs poids de sondage w_k déclarés dans l'instruction WEIGHT.

1.2.2. L'erreur quadratique moyenne estimée du coefficient de régression

La procédure SURVEYREG tient compte du plan de sondage pour estimer l'erreur quadratique moyenne des coefficients de régression, au moyen de la formule générale suivante :

$$\boxed{\hat{V}_{SAS}\left(\hat{\beta}\right) = \left(X_s' W X_s\right)^{-1} G \left(X_s' W X_s\right)^{-1}} \qquad (8)$$

Avec :

$$G = \frac{n-1}{n-p} \sum_{h=1}^{H} \frac{m_h\left(1-f_h\right)}{m_h - 1} \sum_{i=1}^{m_h} \left(e_{hi.} - \bar{e}_{h..}\right)\left(e_{hi.} - \bar{e}_{h..}\right)' \qquad (9)$$

$$e_{hik} = w_{hik} \hat{u}_{hik} X_{hik} \qquad (10)$$

$$e_{hi.} = \sum_{k=1}^{n_{hi}} e_{hik}$$
$$\bar{e}_{h..} = \frac{1}{m_h} \sum_{i=1}^{m_h} e_{hik} \qquad (11)$$

h est l'indice de la strate, i l'indice de l'unité primaire et k l'indice de l'individu dans l'unité primaire. w_{hik} est le poids de sondage de l'unité k, f le taux de sondage. n_{hi} est le nombre d'individus dans la grappe i, m_h est le nombre d'unités primaires dans la strate h.

Dans le cas d'un sondage en grappes ou à plusieurs degrés, f est le taux de sondage du premier degré de tirage.

Pour un sondage non stratifié à un seul degré, les formules (9) à (11) doivent être adaptées en considérant qu'il y a une seule strate (H=1) et que chaque individu constitue une grappe de taille 1 ($n_{hi} = 1$ et $m_h = n$).

X_{hik} est le vecteur (p,1) des valeurs, pour l'individu k, des p variables exogènes X_j du modèle. On reconnaît, dans le produit $m_h \bar{e}_{h..}$, l'expression du vecteur $\hat{t} - \hat{T}\tilde{\beta}$ de la section précédente (§ 1.1.3).

1.2.2.1. Sondage aléatoire simple

Les individus sont sélectionnés dans l'échantillon de taille n avec une probabilité constante égale à $\pi_k = \frac{n}{N}$ et ont pour poids de sondage $w_k = \frac{N}{n}$.

On a alors : $$\boxed{G = N^2 \frac{(1-f)}{n} \frac{(n-1)}{n-p} \left[\mathsf{s}^2\right]} \qquad (12)$$

où $\left[\mathsf{s}^2\right]$ est la matrice (p,p) des variances-covariances empiriques dont les éléments s'écrivent : $s_j^2 = \frac{1}{n-1} \sum_{k \in \mathsf{s}} \left(z_{jk} - \bar{z}_j\right)^2$ et $s_{jq} = \frac{1}{n-1} \sum_{k \in \mathsf{s}} \left(z_{jk} - \bar{z}_j\right)\left(z_{qk} - \bar{z}_q\right)$, la variable z_{jk} étant définie comme précédemment par : $z_{jk} = x_{jk} \hat{u}_k$.

L'expression (12) correspond au développement de (7) dans le cas d'un sondage aléatoire simple, avec une correction par le nombre de degrés de libertés pour le calcul des s_{jq}.

1.2.2.2. Sondage à probabilités proportionnelles à la taille

L'échantillon, non stratifié, est cette fois sélectionné avec des probabilités π_k proportionnelles à un critère de taille des unités. L'expression (9) devient alors :

$$\mathsf{G} = \frac{1-f}{n(n-p)} \mathsf{A} \qquad (13)$$

$$\mathsf{A} = \begin{bmatrix} \cdots & \\ \cdots \sum_{k \in \mathsf{s}} \left(\frac{z_{jk}}{A_k} - \hat{Z}_j\right)\left(\frac{z_{qk}}{A_k} - \hat{Z}_q\right) \cdots \\ \cdots & \end{bmatrix}$$

$$\hat{Z}_j = \sum_{k \in \mathsf{s}} \frac{z_{jk}}{\pi_k} = \sum_{k \in \mathsf{s}} \frac{x_{jk} \hat{u}_k}{\pi_k}$$

$$A_k = \frac{\pi_k}{n}$$

A_k est l'estimation de la probabilité de sélection de l'unité k au i$^{\text{ème}}$ tirage en cas d'indépendance des tirages élémentaires. On reconnaît dans (13) l'estimation de la variance sous l'hypothèse d'un sondage à probabilités inégales avec remise, avec une correction pour population finie $(1-f)$ et une prise en compte du nombre de degrés de liberté $(n-p)$.

1.2.2.3. Sondage en grappes

On suppose que les grappes ne sont pas stratifiées. Rappelons que i est l'indice de la grappe et k celui de l'individu dans la grappe. L'échantillon contient m grappes sélectionnées dans une population qui en comprend M. La matrice G devient :

$$\mathsf{G} = \frac{n-1}{n-p} \frac{(1-f)}{m(m-1)} \mathsf{B} \qquad (14)$$

$$\mathsf{B} = \begin{bmatrix} \cdots \\ \cdots \sum_{i=1}^{m} \left(m \frac{Z_{ji}}{\pi_i} - \hat{Z}_j \right) \left(m \frac{Z_{qi}}{\pi_i} - \hat{Z}_q \right) \cdots \\ \cdots \end{bmatrix}$$

où Z_{ji} est le total de la variable $\mathcal{Z}_j$ dans la grappe i et f le taux de sondage des grappes : $f = \frac{m}{M}$.

Si les grappes sont tirées par sondage aléatoire simple : $\pi_i = \frac{m}{M}$ = taux de sondage des grappes.

On a alors :

$$\mathsf{G} = \frac{n-1}{n-p} M^2 \frac{(1-f)}{m} \begin{bmatrix} s^2_{G1} \cdots s_{(G)1j} \cdots s_{(G)1p} \\ \cdots \quad s^2_{Gj} \quad \cdots \\ s_{(G)p1} \quad \cdots \quad s^2_{Gp} \end{bmatrix} \qquad (15)$$

où : $s_{(G)jq} = \frac{1}{m-1} \sum_{i=1}^{m} \left(Z_{ji} - \frac{\hat{Z}_j}{M} \right) \left(Z_{qi} - \frac{\hat{Z}_q}{M} \right)$ et $s^2_{(G)j} = \frac{1}{m-1} \sum_{i=1}^{m} \left(Z_{ji} - \frac{\hat{Z}_j}{M} \right)^2$

On reconnaît l'expression des variances et covariances estimées entre totaux de grappes des *p* variables $\mathcal{Z}_j$ dans le cas d'un sondage aléatoire simple de grappes.

Si les grappes sont sélectionnées avec des probabilités proportionnelles à leur taille, on a alors : $\frac{m}{\pi_i} = \frac{m}{m \frac{N_i}{N}} = \frac{N}{N_i} \approx \frac{1}{A_i}$, où A_i est la probabilité de sélection de la grappe *i* à chaque tirage élémentaire lorsque ceux-ci sont indépendants. On a alors dans (14) l'expression des variances-covariances propres au sondage à probabilités inégales avec remise, avec une correction pour population finie $(1-f)$ et la prise en compte du nombre de degrés de liberté de la régression $(n-p)$.

1.2.2.4. Sondage à deux degrés

Le développement des formules (9) à (11) de la procédure SURVEYREG nous donne l'expression suivante de la matrice G :

$$G = \frac{n-1}{n-p} \frac{1-f}{m} \frac{1}{m-1} \mathsf{C}$$

$$\mathsf{C} = \begin{bmatrix} \cdots \\ \cdots \left(\sum_{i=1}^{m} \left(m \frac{\hat{Z}_{ji}}{\pi_i} - \hat{Z}_j \right) \left(m \frac{\hat{Z}_{li}}{\pi_i} - \hat{Z}_l \right) \right) \cdots \\ \cdots \end{bmatrix} \qquad (16)$$

où m est le nombre d'unités primaires sélectionnées, f le taux de sondage au premier degré de tirage et π_i la probabilité d'inclusion de l'UP i au premier degré. $\hat{Z}_{ji}$ est l'estimateur Horvitz-Thomson du total de Z_j dans l'unité primaire i, $\hat{Z}_j$ l'estimateur du total dans la population. La variance est estimée sous l'hypothèse d'un tirage avec remise des unités primaires. On a donc une estimation de la variance due au premier degré de sondage seul.

1.2.2.5. Sondage stratifié

Quel que soit le mode de tirage à l'intérieur des strates, la variance est alors égale à la somme des variances de strates. La variance interne à une strate est calculée selon les formules précédentes en fonction du plan de sondage adopté dans la strate. La procédure fait l'hypothèse d'un plan de sondage identique dans toutes les strates.

1.2.3. Le tableau d'analyse de la variance

Le tableau ANOVA fourni par SURVEYREG est conforme aux formules indiquées dans la section 1.1.4, dans lesquelles $\hat{\bar{Y}}$ est l'estimateur de la moyenne dans la population :

$$\hat{\bar{Y}} = \frac{1}{\hat{N}} \sum_{k \in s} \frac{y_k}{\pi_k} \qquad \text{avec :} \quad \hat{N} = \sum_{k \in s} \frac{1}{\pi_k}.$$

La somme totale des carrés est calculée comme suit :

$$SCT = \sum_{k \in s} \frac{\left(y_k - \hat{\bar{Y}}\right)^2}{\pi_k} \quad \text{lorsque la régression comprend un terme constant}$$

$$SCT = \sum_{k \in s} \frac{{y_k}^2}{\pi_k} \quad \text{lorsque la régression ne comprend pas de terme constant}$$

La somme des carrés des résidus est une somme pondérée : $SCE = \sum_{k \in s} \frac{\hat{u}_k^{\,2}}{\pi_k}$.

Le coefficient de détermination est le rapport : $\left(1 - SCE / SCT\right)$.

1.2.4. Le test de Student de nullité d'un coefficient de régression

La fonction du test est le rapport : $\hat{T} = \frac{\hat{b}_j}{\hat{\sigma}_{\hat{b}_j}}$ dans lequel le dénominateur est la racine carrée de la variance estimée du coefficient $\hat{b}_j$ en fonction du plan de sondage, avec les approximations indiquées plus haut.

1.2.5. Différence entre les procédures REG et SURVEYREG

On peut réaliser une régression pondérée avec la procédure REG. Lorsqu'on déclare dans les deux procédures le poids de sondage de l'observation en variable de pondération (instruction WEIGHT), on obtient les résultats suivants :

- le tableau d'analyse de la variance est le même
- le coefficient de détermination est identique
- le vecteur $\hat{b}$ des coefficients de régression estimés est identique
- les écarts-types estimés des coefficients $\hat{b}$ et les t-values associées diffèrent, puisque dans la procédure REG, ils sont calculés sans tenir compte du plan de sondage.

La procédure REG estime la matrice des variances-covariances des coefficients de régression $\hat{b}_j$ de la façon suivante :

- sans pondération : $\hat{V}\left(\hat{\mathbf{b}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\tilde{\sigma}^2_{\hat{u}}$, où $\tilde{\sigma}^2_{\hat{u}} = \dfrac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-p}$.
- en présence d'une instruction WEIGHT : $\hat{V}\left(\hat{\mathbf{b}}\right) = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\tilde{\sigma}^2_{\hat{u}}$, avec $\mathbf{W} = Diag\left(w_k\right)$.

Il s'ensuit que les tests de Student de nullité des coefficients de régression, c'est-à-dire l'appréciation du caractère significatif ou non des variables explicatives, peuvent conduire à des résultats différents entre les deux procédures.

1.3. Exemple de régression avec la procédure SURVEYREG

L'économie du milieu rural est fortement différenciée selon la densité des relations qui lient les communes rurales aux villes proches. La proportion de la population travaillant dans une autre commune que celle où elle réside est une mesure synthétique de cette inter-pénétration. On s'intéresse aux facteurs qui influencent l'importance de ces migrations quotidiennes de travail.

La population est constituée des 2306 cantons ruraux. Le pourcentage de la population active résidant dans le canton et exerçant un emploi dans une autre commune que celle où elle réside est la variable expliquée (*migrants*). On estime une relation entre cette variable et les facteurs suivants :

- nombre d'agriculteurs, exprimé en pourcentage de la population active résidente
- nombre de cadres, exprimé en pourcentage de la population active résidente
- nombre de personnes de 60 ans ou plus, exprimé en pourcentage de la population résidente
- part des logements de 4 pièces ou plus dans le parc de logements.

1000 échantillons de 120 cantons ont été sélectionnés selon trois plans de sondage : sondage aléatoire simple, sondage à probabilités proportionnelles à la population, sondage stratifié selon le poids des jeunes et celui des agriculteurs dans la population avec des probabilités proportionnelles à la population.

Le modèle a été estimé dans la population de référence avec la procédure REG, puis dans chaque échantillon successivement avec la procédure REG et avec la procédure SURVEYREG utilisant les poids de sondage en pondération dans les deux cas. Les résultats sont résumés par les médianes des paramètres obtenus.

1.3.1. Sondage aléatoire simple

Paramètres	**Population de référence**	**REG pondéré**	**SURVEYREG**
R2 du modèle	0,766	0,774	0,774
Coefficients de régression :			
• Terme constant	52,164	52,591	52,591
• Agriculteurs	-0,558	-0,550	-0,550
• Cadres	0,459	0,465	0,465
• Plus de 60 ans	-0,646	-0,654	-0,654
• Logements de 4 pièces ou +	0,370	0,368	0,368
Ecart-type empirique[1] des coefficients de régression :			
• Terme constant	///	8,520	8,520
• Agriculteurs	///	0,100	0,100
• Cadres	///	0,265	0,265
• Plus de 60 ans	///	0,174	0,174
• Logements de 4 pièces ou +	///	0,086	0,086
Ecarts-types estimés[2] des coefficients de régression :			
• Terme constant	1,303	5,794	7,419
• Agriculteurs	0,022	0,096	0,095
• Cadres	0,057	0,260	0,242
• Plus de 60 ans	0,032	0,144	0,160
• Logements de 4 pièces ou +	0,013	0,056	0,075
T-values estimées[3] des coefficients de régression :			
• Terme constant	40,05	9,08	7,00
• Agriculteurs	-25,88	-5,77	-5,84
• Cadres	8,07	1,86	2,04
• Plus de 60 ans	-19,85	-4,55	-4,12
• Logements de 4 pièces ou +	29,17	6,53	4,67

[1] Ecart-type des coefficients estimés dans la distribution des 1000 échantillons : $\hat{s}_{j\text{emp}} = \sqrt{\frac{1}{999}\sum_{s=1}^{1000}\left(\hat{b}_{js} - \bar{\hat{b}}_j\right)^2}$,

avec $\bar{\hat{b}}_j = \frac{1}{1000}\sum_{s=1}^{1000}\hat{b}_{js}$

[2] Médiane des estimateurs de l'écart-type par échantillon.

[3] Il s'agit de la médiane des t-values obtenues dans les 1000 échantillons simulés. Par conséquent, elle n'est pas exactement égale au rapport entre la médiane du coefficient de régression et la médiane de son écart-type estimé, présentés dans ce tableau.

1.3.2. Sondage à probabilités proportionnelles à la taille

Paramètres	**Population de référence**	**REG pondéré**	**SURVEYREG**
R2 du modèle	0,766	0,782	0 ,782
Coefficients de régression :			
• Terme constant	52,164	55,346	55,346
• Agriculteurs	-0,558	-0,562	-0,562
• Cadres	0,459	0,444	0,444
• Plus de 60 ans	-0,646	-0,693	-0,693
• Logements de 4 pièces ou +	0,370	0,339	0,339
Ecart-type empirique[4] des coefficients de régression :			
• Terme constant	///	11,005	11,005
• Agriculteurs	///	0,115	0,115
• Cadres	///	0,285	0,285
• Plus de 60 ans	///	0,214	0,214
• Logements de 4 pièces ou +		0,112	0,112
Ecarts-types estimés[5] des coefficients de régression :			
• Terme constant	1,303	5,848	8,276
• Agriculteurs	0,022	0,094	0,106
• Cadres	0,057	0,254	0,254
• Plus de 60 ans	0,032	0,145	0,179
• Logements de 4 pièces ou +	0,013	0,057	0,085
T-values[6] des coefficients de régression :			
• Terme constant	40,05	9,36	6,59
• Agriculteurs	-25,88	-5,95	-5,28
• Cadres	8,07	1,81	1,82
• Plus de 60 ans	-19,85	-4,71	-3,79
• Logements de 4 pièces ou +	29,17	5,84	3,79

[4] Voir note 1 du § 1.3.1
[5] Voir note 2 du § 1.3.1
[6] Voir note 3 du § 1.3.1

1.3.3. Sondage stratifié à probabilités proportionnelles à la taille

Paramètres	**Population de référence**	**REG pondéré**	**SURVEYREG**
R2 du modèle	0,766	0,781	0,781
Coefficients de régression :			
• Terme constant	52,164	55,039	55,039
• Agriculteurs	-0,558	-0,559	-0,559
• Cadres	0,459	0,451	0,451
• Plus de 60 ans	-0,646	-0,701	-0,701
• Logements de 4 pièces ou +	0,370	0,340	0,340
Ecart-type empirique[7] des coefficients de régression :			
• Terme constant	///	11,555	11,555
• Agriculteurs	///	0,111	0,111
• Cadres	///	0,293	0,293
• Plus de 60 ans	///	0,216	0,216
• Logements de 4 pièces ou +		0,117	0,117
Ecarts-types estimés[8] des coefficients de régression :			
• Terme constant	1,303	5,873	8,106
• Agriculteurs	0,022	0,095	0,103
• Cadres	0,057	0,259	0,255
• Plus de 60 ans	0,032	0,144	0,174
• Logements de 4 pièces ou +	0,013	0,057	0,084
T-values[9] des coefficients de régression :			
• Terme constant	40,05	9,37	6,80
• Agriculteurs	-25,88	-5,94	-5,44
• Cadres	8,07	1,77	1,85
• Plus de 60 ans	-19,85	-4,84	-3,93
• Logements de 4 pièces ou +	29,17	5,90	3,77

L'écart-type empirique des 1000 coefficients de régression estimés pour chaque variable du modèle reflète au plus près la variance vraie du paramètre. On voit qu'avec la procédure SURVEYREG, l'estimateur de la variance approche mieux la variance empirique que celui fourni par la procédure REG pondérée. Il s'ensuit des t-values différentes qui peuvent modifier le jugement porté sur le caractère significatif d'un facteur. C'est le cas ici pour la variable « cadres » dans le sondage aléatoire simple.

[7] Voir note 1 du § 1.3.1
[8] Voir note 2 du § 1.3.1
[9] Voir note 3 du § 1.3.1

2. L'analyse de la variance

2.1. Rappels théoriques

La variable d'intérêt Y est une variable numérique continue, et l'on cherche à mesurer l'influence de q caractères qualitatifs X^r à p_r modalités sur le niveau de Y. L'on souhaite vérifier si les différences entre les moyennes de Y selon les classes de population définies par X^r sont significatives.

Les notations sont les mêmes que précédemment, mais ici, $\mathcal{X}_j^r$ est la variable indicatrice prenant la valeur 1 si l'individu présente la modalité j du caractère X^r et 0 sinon. Il y a donc p_r variables indicatrices $\mathcal{X}_j^r$.

Le vecteur ligne $\mathbf{X}_k' = \left[...X_{1k}^r...X_{jk}^r...X_{pk}^r...\right]$ des valeurs observées pour l'individu k comprend q coordonnées égales à 1 et $\sum_{r=1}^{q}\left(p_r - 1\right)$ coordonnées égales à 0. $\mathbf{U}_j^r$ est la sous-population de taille N_j^r des individus vérifiant le caractère j de X^r. $\mathbf{s}_j^r$ est le sous-échantillon de taille n_j^r (aléatoire) qui contient les individus vérifiant le caractère j de X^r.

2.1.1. L'estimation des paramètres

Avec un facteur, le modèle s'écrit :

$$Y_k = \sum_{j=1}^{p} \mu_j X_{jk} + U_k$$

$$\mathbf{Y} = \mathbf{X}\theta + \mathbf{u} \qquad (17)$$

où u représente l'ensemble des facteurs non pris en compte.

Avec deux facteurs X et Z ayant respectivement p et q modalités, le modèle se formalise :

$$\mathbf{Y} = \mathbf{X}\alpha + \mathbf{Z}\beta + \mathbf{T}\gamma + \mathbf{u}$$

La variable $\mathcal{T}_{ij}$ est l'indicatrice associée au croisement des critères i et j de X et Z . T représente l'interaction entre les facteurs X et Z.

Le vecteur θ des coefficients de régression est estimé dans l'échantillon par les MCO pondérés comme dans le cas de la régression linéaire :

$$\hat{\theta} = \left(\mathbf{X}_s' \mathbf{W} \mathbf{X}_s\right)^{-} \mathbf{X}_s' \mathbf{W} \mathbf{Y}_s \qquad (18)$$

Lorsque le modèle comprend plusieurs facteurs, ou lorsqu'il comprend un facteur et un terme constant, les vecteurs de $\mathbf{X}_s$ sont colinéaires et la matrice $\mathbf{X}_s' \mathbf{W} \mathbf{X}_s$ n'est pas inversible. On lui substitue dans (18) une inverse généralisée, ce qui revient à imposer des contraintes identifiantes, par exemple en annulant le paramètre de la dernière modalité de chaque facteur à partir du deuxième.

L'erreur quadratique moyenne de θ s'estime selon les mêmes relations (6) et (7) du § 1.1.3.

2.1.2. Test d'une relation linéaire sur les paramètres

L'hypothèse d'une relation linéaire entre les coefficients μ_j est formalisée par $H_0 : c'\theta = 0$, où c est un vecteur de coefficients.

Dans la population, $\tilde{\theta}$ étant l'estimateur des MCO de θ, la fonction $\varphi = c'\theta$ est estimée sans biais par : $\tilde{\varphi} = c'\tilde{\theta}$.

Sa variance : $V(\tilde{\varphi}) = c'V(\tilde{\theta})c = \sigma_u^2 c'(X'X)^{-1}c$ (19)

est estimée par : $\tilde{V}(\tilde{\varphi}) = c'\tilde{V}(\tilde{\theta})c = \tilde{\sigma}_u^2 c'(X'X)^{-1}c = \frac{\tilde{u}'\tilde{u}}{N-p} c'(X'X)^{-1}c$ (20)

Sous l'hypothèse de normalité et d'homoscédasticité des aléas, $\frac{\tilde{u}'\tilde{u}}{\sigma_u^2} = \frac{(N-p)\tilde{\sigma}_u^2}{\sigma_u^2}$ suit une loi χ^2_{N-p} et $(c'\tilde{\theta} - c'\theta)$ suit une loi normale $\mathcal{N}(0, c'V(\tilde{\theta})c)$.

D'où : $(c'\tilde{\theta} - c'\theta)[c'V(\tilde{\theta})c]^{-1}(c'\tilde{\theta} - c'\theta)$ suit une loi $\chi^2_{p_0}$ où p_0 est le nombre de paramètres indépendants sous Ho.

On a donc : $F = \frac{\frac{1}{p_0}(c'\tilde{\theta} - c'\theta)[c'V(\tilde{\theta})c]^{-1}(c'\tilde{\theta} - c'\theta)}{\frac{1}{(N-p)}\frac{\tilde{u}'\tilde{u}}{\sigma_u^2}} = \frac{\frac{1}{p_0}(c'\tilde{\theta} - c'\theta)[c'(X'X)^{-1}c]^{-1}(c'\tilde{\theta} - c'\theta)}{\frac{1}{(N-p)}\tilde{u}'\tilde{u}}$

suit une loi $\mathcal{F}(p_0, N-p)$.

Sous H_0 : $F = \frac{\frac{1}{p_o}(c'\tilde{\theta})[c'(X'X)^{-1}c]^{-1}c'\tilde{\theta}}{\frac{1}{N-p}\tilde{u}'\tilde{u}} = \frac{1}{p_0}(c'\tilde{\theta})'[c'\tilde{V}(\tilde{\theta})c]^{-1}(c'\tilde{\theta})$ (21)

suit une loi de Fisher : $\mathcal{F}(p_0, N-p)$ où p_0 est le nombre de paramètres indépendants sous Ho, et p le rang de la matrice $X'X$ dans le modèle complet. Lorsque la valeur observée de F est supérieure à $\mathcal{F}(p_0, N-p, 1-\alpha)$, où $\mathcal{F}(p_0, N-p, 1-\alpha)$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher de paramètres $(p_0, N-p)$, on rejette H_0.

Dans l'échantillon, on teste l'hypothèse $H_0 : c'\theta = 0$ au moyen de la fonction (21) dans laquelle on substitue aux grandeurs inconnues leurs estimateurs :

- $\tilde{\theta}$ est remplacé par son estimateur $\hat{\theta}$
- $\tilde{V}(\tilde{\theta})$ est remplacé par l'estimateur $\hat{V}(\hat{\theta})$

La relation (21) devient : $\hat{F} = \frac{1}{p_0}(c'\hat{\theta})'[c'\hat{V}(\hat{\theta})c]^{-1}(c'\hat{\theta})$ qui suit approximativement une loi de Fisher à $(p_0, n-p)$ degrés de liberté, où p_0 est le nombre de paramètres indépendants sous H_0.

2.1.3. Cas particulier du modèle à un facteur

L'estimateur $\hat{\mu}_j$ du coefficient de régression est alors égal à l'estimateur Horvitz-Thomson $\hat{\bar{Y}}_j$ de la moyenne de Y dans la sous-population $\mathrm{U_j}$. Sa variance est donc celle d'une moyenne mesurée à partir d'un sous-échantillon $\mathrm{s_j}$ dont la taille n_j est elle-même aléatoire.

Lorsque la taille N_j du domaine $\mathrm{U_j}$ est connue, la moyenne et la variance de Y dans le domaine $\mathrm{U_j}$ sont estimées par :

$$\hat{\bar{Y}}_j = \frac{\hat{Y}_j}{N_j} = \frac{1}{N_j}\sum_{k\in \mathrm{s}_j}\frac{y_k}{\pi_k} \quad (22) \qquad \hat{V}\left(\hat{\bar{Y}}_j\right) = \frac{\hat{V}\left(\hat{Y}_j\right)}{N_j^2} = \frac{1}{N_j^2}\sum_{k\in \mathrm{s}_j}\sum_{l\in \mathrm{s}_j}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \quad (23)$$

Lorsque la taille du domaine $\mathrm{U_j}$ est inconnue, on lui substitue son estimateur : $\hat{N}_j = \sum_{k\in \mathrm{s}_j}\frac{1}{\pi_k}$ dans l'expression (22). La variance de l'estimateur de la moyenne dans le domaine est alors obtenue après linéarisation de $\hat{\bar{Y}}_j$ par la méthode de Taylor, ce qui donne :

$$\hat{V}\left(\hat{\bar{Y}}_j\right) = \frac{1}{\hat{N}_j^2}\hat{V}\left(\sum_{k\in \mathrm{s}}\frac{x_{jk}\left(y_k - \hat{\bar{Y}}_j\right)}{\pi_k}\right) = \frac{1}{\hat{N}_j^2}\sum_{k\in \mathrm{s}_j}\sum_{l\in \mathrm{s}_j}\frac{\Delta_{kl}}{\pi_{kl}}\frac{\left(y_k - \hat{\bar{Y}}_j\right)}{\pi_k}\frac{\left(y_l - \hat{\bar{Y}}_j\right)}{\pi_l} \quad (24)$$

Dans le modèle (17) ajusté sur l'échantillon s, la variance du paramètre μ_j est estimée par :

$$\hat{V}\left(\hat{\mu}_j\right) = \frac{n-1}{n-p}\hat{V}\left(\hat{\bar{Y}}_j\right) \quad (25)$$

2.2. Analyse de la variance avec SURVEYREG

2.2.1. L'estimation des coefficients et de leur précision

La procédure SURVEYREG utilise les mêmes formules générales d'estimation que dans le cas de la régression linéaire. Dans le cas du modèle à un facteur, elle nous donne des coefficients de régression égaux aux estimateurs $\hat{\bar{Y}}_j$ des moyennes de classes. La variance des coefficients de régression est estimée au moyen des formules (8) à (11).

Dans le cas du modèle à un facteur, la matrice $\left(\mathsf{X}_\mathsf{s}'\mathsf{W}\mathsf{X}_\mathsf{s}\right)$ est une matrice diagonale de termes : $\sum_{k\in \mathrm{s}}\frac{x_{jk}}{\pi_k} = \sum_{k\in \mathrm{s_j}}\frac{1}{\pi_k} = \hat{N}_j$ où $\hat{N}_j$ est l'estimateur de l'effectif de la sous-population $\mathrm{U_j}$. La matrice inverse $\left(\mathsf{X}_\mathsf{s}'\mathsf{W}\mathsf{X}_\mathsf{s}\right)^{-1}$ est par conséquent une matrice diagonale de termes : $1/\hat{N}_j$.

L'expression (8) prend alors la forme générale suivante :

$$\hat{V}_{SAS}\left(\hat{\mu}_j\right)=\frac{n-1}{n-p}\frac{1}{\hat{N}_j^2}\hat{V}\left[\sum_{k\in s}\frac{z_{jk}}{\pi_k}\right]=\frac{n-1}{n-p}\frac{1}{\hat{N}_j^2}\hat{V}\left[\sum_{k\in s}\frac{x_{jk}\left(y_k-\hat{\bar{Y}}_j\right)}{\pi_k}\right] \qquad (26)$$

avec : $z_{jk}=x_{jk}\hat{u}_k=x_{jk}\left(y_k-\hat{\bar{Y}}_j\right)$ (27)

La variable Z définie en (27) correspond à la variable auxiliaire utilisée pour l'approximation de la variance d'un ratio par linéarisation, lorsque X_j est une indicatrice d'appartenance à une sous-population. On retrouve, à partir des formules (8) à (11), l'expression générale (24) de la variance estimée d'une moyenne dans un domaine, sous le plan de sondage adopté, avec une correction pour population finie $(1-f)$ optionnelle, et une correction par le nombre de degrés de liberté du modèle : $\frac{n-1}{n-p}$.

Lorsque le plan de sondage est à probabilités proportionnelles à la taille, l'hypothèse est faite d'un sondage avec remise.

Lorsque le sondage est à plusieurs degrés, l'hypothèse est faite d'un sondage avec remise au premier degré, ce qui conduit à une estimation de la variance qui ne tient pas compte de la variance due au deuxième tirage ni aux tirages successifs.

2.2.2. Les contraintes identifiantes dans le modèle à un facteur avec constante

En présence d'un terme constant, le modèle (17) devient : $\mathbf{Y}=\mathbf{X}\theta+\mathbf{m}+\mathbf{u}$ où $\mathbf{m}$ est un vecteur dont les coordonnées sont une constante. Ce qu'on peut encore écrire :

$$\mathbf{Y}=\mathbf{Z}\mathbf{T}+\mathbf{u} \qquad (28) \qquad \text{où :} \quad \mathbf{Z}=\begin{bmatrix}1 & X_{11}\ldots X_{j1}\ldots X_{p1}\\ 1 & \ldots \quad X_{jk} \quad \ldots\\ 1 & X_{1n}\ldots X_{jn}\ldots X_{pn}\end{bmatrix} \qquad \mathbf{T}=\begin{bmatrix}m\\ \mu_1\\ \ldots\\ \mu_p\end{bmatrix}$$

Le modèle (28) n'est pas identifiable car la matrice Z n'est pas régulière. Pour résoudre les équations normales, la procédure SURVEYREG contraint à 0 le dernier coefficient : $\mu_p=0$.

Les coefficients de régression estiment alors les différences respectives entre la moyenne de Y dans une classe définie par la modalité *j* du facteur X et la moyenne dans la sous-population U_p définie par la dernière modalité *p* de X : $\hat{\mu}_j=\hat{\bar{Y}}_j-\hat{\bar{Y}}_p$.

2.2.3. Les contraintes identifiantes dans le modèle à plusieurs facteurs

La procédure calcule l'inverse généralisée de la matrice $\mathbf{X'WX}$, en annulant les paramètres des modalités représentées dans la matrice X par un vecteur égal à une combinaison linéaire des vecteurs précédents.

Dans un modèle sans interaction, cela conduit à annuler le paramètre de la dernière modalité de chaque effet si le modèle contient un terme constant, de chaque effet à partir du deuxième en l'absence de terme contant. Si l'on a spécifié le modèle : $\mathsf{Y} = \mathsf{X}\ \mathsf{Z}$, le facteur X ayant p modalités et le facteur Z q modalités, la procédure SURVEYREG annule la p^{ième} modalité de X et la q^{ième} modalité de Z par défaut, la q^{ième} modalité de Z seulement si l'on a demandé la suppression du terme constant.

2.2.4. Test des effets

Le tableau d'analyse de la variance fourni par la procédure SURVEYREG (en version 8) ne tient pas compte du plan de sondage. Il ne diffère donc pas de celui fourni par la procédure GLM utilisant la même variable de pondération.

En l'absence de terme constant, la procédure SURVEYREG calcule le coefficient de détermination comme : $1-\left(\sum_{k\in s}\frac{\hat{u}_k^2}{\pi_k}/\sum_{k\in s}\frac{y_k^2}{\pi_k}\right)$, contrairement à la procédure GLM qui reproduit la valeur du R^2 correspondant au modèle avec terme constant.

La procédure SURVEYREG teste l'efficacité des effets du modèle au moyen des tests de type III. Les résultats sont donc indifférents à l'ordre dans lequel sont spécifiés les paramètres dans l'instruction MODEL. Les calculs mis en œuvre pour ces tests utilisent la variance estimée sous le plan de sondage, tel qu'il est spécifié au travers de la valeur des pondérations indiquées dans l'instruction WEIGHT et des instructions éventuelles concernant la stratification et les grappes. La variance des coefficients de régression est estimée selon les conventions décrites plus haut.

Le test d'une hypothèse linéaire $\mathsf{L}\theta = 0$ s'appuie sur la statistique : $F_W = \frac{\left(\mathsf{L}\hat{\theta}\right)'\left(\mathsf{L}'\hat{\mathsf{V}}\mathsf{L}\right)^{-}\left(\mathsf{L}\hat{\theta}\right)}{r}$.

$\hat{\theta}$ est le vecteur des coefficients de régression estimés, $\hat{\mathsf{V}}$ la matrice de leurs variances-covariances estimées, L une matrice associée à la relation linéaire testée, éventuellement multiple, et telle que la fonction $\mathsf{L}\theta$ soit estimable, r est le rang de la matrice L.

F_W est comparé au quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à $(r, m-H)$ degrés de liberté, où m est le nombre d'unités primaires de l'échantillon et H le nombre de strates. Avec un sondage à un degré non stratifié, on a (n-1) degrés de liberté.

2.3. Exemple d'analyse de la variance avec SURVEYREG

L'enquête santé de 2001 (partie variable de l'enquête permanente sur les conditions de vie des ménages) a servi de population de référence aux exercices de simulation ci-dessous. 5194 individus âgés de 15 ans ou plus avaient répondu à cette enquête. On a sélectionné 100 échantillons de taille 500 dans ce référentiel, pour modéliser le taux de recours au médecin généraliste par individu.

La variable expliquée (*GENPC*) est le nombre de fois où la personne a consulté un généraliste au cours des 12 mois précédant l'enquête.

Trois critères exercent une influence significative sur la fréquence des consultations :

- l'âge
- le sexe
- le degré de couverture sociale.

On a retenu trois tranches d'âge :

- moins de 50 ans
- 50 à 79 ans
- 80 ans ou plus

et trois niveaux de couverture sociale (variable *SECUR*) :

- sécurité sociale et pas de mutuelle
- sécurité sociale et mutuelle
- couverture maladie universelle (CMU).

Ce modèle à trois facteurs sans interaction est estimé dans chaque échantillon, successivement avec la procédure GLM utilisant l'option WEIGHT et avec la procédure SURVEYREG. Les observations sont pondérées par leur poids de sondage dans les deux procédures. Les résultats sont résumés par les médianes des 100 « F-value » et des 100 « T-value » obtenues.

Un exemple de programme de régression avec la procédure SURVEYREG :

```
PROC SURVEYREG DATA=echant RATE=plan ;
    STRATA sante ;
     WEIGHT poids ;
     CLASS age sexe secur ;
     MODEL genpc=age sexe secur / SOLUTION;
RUN ;
```

Dans le programme ci-dessus, l'échantillon est stratifié selon la variable « sante » et sélectionné par sondage aléatoire simple dans les strates. Les taux de sondage par strate sont stockés dans la table « plan » et spécifiés dans une option RATE pour l'intégration du facteur $(1-f_h)$ à l'estimation de la variance par strate.

2.3.1. Sondage aléatoire simple

Paramètres	**Population de référence**	**GLM pondéré**	**SURVEYREG**
F-value du modèle	966,95	97,00	103,24
F-value des effets (type III)			
• Age	217,38	22,43	24,81
• Sexe	67,41	5,93	6,96
• Secur	88,76	9,90	6,44
Ecarts-types des coefficients de régression :			
• Moins de 50 ans	0,2665	0,8648	0,7146
• 50 à 79 ans	0,2750	0,8858	0,7699
• 80 ans et plus	0,3712	1,1955	1,1188
• Femmes	0,1320	0,4227	0,3905
• couverture maladie universelle (CMU)	0,3392	1,0966	1,2606
• sécurité sociale et mutuelle	0,2633	0,8529	0,7070
T-values des coefficients de régression :			
• Moins de 50 ans	7,36	2,04	2,28
• 50 à 79 ans	15,28	4,62	4,69
• 80 ans et plus	17,68	5,19	5,43
• Femmes	8,21	2,43	2,64
• couverture maladie universelle (CMU)	10,85	3,49	3,15
• sécurité sociale et mutuelle	2,45	1,07	1,16

Les modalités « hommes » et « sécurité sociale sans mutuelle » ont été mises à zéro.

2.3.2. Sondage stratifié avec sondage aléatoire simple dans les strates

Le tirage est stratifié selon l'état de santé ressenti par l'individu : bon, moyen ou mauvais. On a donc trois strates.

Paramètres	**Population de référence**	**GLM pondéré**	**SURVEYREG**
F-value du modèle	966,95	96,33	118,87
F-value des effets (type III)			
• Age	217,38	22,04	23,87
• Sexe	67,41	7,35	8,34
• Secur	88,76	8,84	6,43
Ecarts-types des coefficients de régression :			
• Moins de 50 ans	0,2665	0,8642	0,7848
• 50 à 79 ans	0,2750	0,8925	0,8272
• 80 ans et plus	0,3712	1,2106	1,1535
• Femmes	0,1320	0,4242	0,3990
• couverture maladie universelle (CMU)	0,3392	1,1063	1,3212
• sécurité sociale et mutuelle	0,2633	0,8526	0,7656
T-values des coefficients de régression :			
• Moins de 50 ans	7,36	2,21	2,26
• 50 à 79 ans	15,28	4,50	4,55
• 80 ans et plus	17,68	5,36	5,35
• Femmes	8,21	2,71	2,89
• couverture maladie universelle (CMU)	10,85	3,45	2,98
• sécurité sociale et mutuelle	2,45	1,07	1,15

2.3.3. Sondage à deux degrés

Au premier degré, 5 régions sont sélectionnées parmi 20 avec des probabilités proportionnelles à leur effectif dans la population de référence. Au second degré, on tire 100 individus dans chaque région sélectionnée par sondage aléatoire simple. On a donc des échantillons de taille 500.

Pour obtenir 100 échantillons distincts, on a sélectionné 10 échantillons de régions, et 10 échantillons d'individus dans chaque échantillon d'unités primaires.

Paramètres	**Population de référence**	**GLM pondéré**	**SURVEYREG**
F-value du modèle	966,95	97,43	141,73
F-value des effets (type III)			
• Age	217,38	21,29	30,22
• Sexe	67,41	6,80	9,85
• Secur	88,76	8,40	13,48
Ecarts-types des coefficients de régression :			
• Moins de 50 ans	0,2665	0,8497	0,5457
• 50 à 79 ans	0,2750	0,8781	0,7948
• 80 ans et plus	0,3712	1,2179	0,9840
• Femmes	0,1320	0,4185	0,3420
• couverture maladie universelle (CMU)	0,3392	1,0892	0,9274
• sécurité sociale et mutuelle	0,2633	0,8377	0,5841
T-values des coefficients de régression :			
• Moins de 50 ans	7,36	2,28	3,28
• 50 à 79 ans	15,28	4,96	5,36
• 80 ans et plus	17,68	5,31	6,07
• Femmes	8,21	2,61	3,14
• couverture maladie universelle (CMU)	10,85	3,20	3,82
• sécurité sociale et mutuelle	2,45	0,71	1,09

La prise en compte du plan de sondage dans la procédure SURVEYREG induit ici des écarts-types plus faibles, donc des « F-values » et des « T-values » plus élevées. Ceci tient notamment, dans le sondage aléatoire simple et le sondage stratifié, à l'application du facteur correcteur (1-*f*), important dans le cas présent. Les résultats obtenus avec le sondage à deux degrés sont plus ambigus. Dans la mesure où la procédure ne calcule que la variance due au premier degré de sondage, on peut craindre une sous-estimation de la variance réelle lorsque l'hypothèse « avec remise » est moins défendable. C'est le cas avec un fort taux de sondage au premier degré.

3. La régression logistique

3.1. Le modèle dichotomique : rappels théoriques

On cherche à expliquer un phénomène de nature dichotomique par un ensemble de facteurs. Par exemple, on souhaite mesurer les caractéristiques socio-économiques qui influencent la non-réponse dans une enquête auprès des ménages. La variable expliquée est catégorielle à valeurs 0 ou 1, les variables exogènes peuvent être numériques ou catégorielles. La régression linéaire est inopérante dans ce cas et l'on passe par l'intermédiaire d'une variable latente pour modéliser le mécanisme observé.

Soient Y la variable dichotomique à expliquer, Y^* une variable latente, $X_1,...X_J$ J variables exogènes observées, $b_1,...b_J$ des coefficients.

Y^* est une variable continue définie par : $Y^* \geq seuil \Rightarrow Y=1$

$$Y^* < seuil \Rightarrow Y=0$$

Par convention, on choisit un seuil égal à 0, et l'on modélise la probabilité de dépassement de ce seuil, en postulant une relation linéaire entre la variable latente inobservée et les facteurs exogènes connus. Avec les mêmes notations que précédemment :

$$Y_k^* = \sum_{j=1}^{J} b_j X_{jk} + u_k = \mathbf{b}'\mathbf{X}_k + u_k \qquad (29)$$

$$p_k = \text{Prob}\{Y_k = 1\} = \text{Prob}\{Y_k^* \geq 0\} = \text{Prob}\{\mathbf{b}'\mathbf{X}_k + u_k \geq 0\} = F(\mathbf{b}'\mathbf{X}_k) \qquad (30)$$

où u représente l'ensemble des facteurs non pris en compte par le modèle et F la fonction de répartition de $-u$.

Le modèle LOGIT utilise la fonction de répartition de la loi logistique :

$$F(t) = \frac{1}{1+\exp(-t)} \qquad (31)$$

d'où : $$p_k = \text{Prob}\{Y_k = 1\} = F(\mathbf{b}'\mathbf{X}_k) = \frac{1}{1+\exp(-\mathbf{b}'\mathbf{X}_k)} \qquad (32)$$

Le modèle PROBIT utilise la fonction de répartition de la loi normale.

Dans ce qui suit, on se limitera à la présentation du modèle LOGIT.

3.1.1. L'estimation des paramètres

Le vecteur des paramètres du modèle est estimé par la méthode du maximum de vraisemblance. **Dans la population**, l'estimateur $\tilde{\mathbf{b}}$ est solution des équations $\frac{\delta Log(L)}{\delta b_j}=0$ où L est la vraisemblance.

$$L = \prod_{k \in \text{Pop}} F(b' X_k)^{Y_k} \left[1 - F(b' X_k)\right]^{1-Y_k}$$

$$Log(L) = \sum_{k \in Pop} \left(Y_k Log\left[F(b' X_k)\right] + (1 - Y_k) Log\left[1 - F(b' X_k)\right] \right) \quad (33)$$

$$\frac{\delta Log(L)}{\delta b} = \sum_{k \in Pop} \frac{Y_k - F(\tilde{b}' X_k)}{F(\tilde{b}' X_k)(1 - F(\tilde{b}' X_k))} f(\tilde{b}' X_k) X_k = \sum_{k \in Pop} \left[Y_k - F(\tilde{b}' X_k)\right] X_k = 0 \quad (34)$$

où *f* est la dérivée de la fonction *F*, c'est-à-dire la densité de probabilité.

Dans un échantillon, π_k étant la probabilité d'inclusion de l'unité *k*, la log-vraisemblance (33) est estimée par l'estimateur Horvitz-Thomson, et les coefficients de régression sont solutions des équations :

$$\frac{\delta L\hat{o}g(L)}{\delta b} = 0 \quad \Leftrightarrow \quad \sum_{k \in s} \frac{1}{\pi_k} \left[y_k - F(\hat{b}' x_k)\right] x_k = 0 \quad (35)$$

3.1.2. La précision de l'estimateur

Soit : $$l(b) = \sum_{k \in Pop} \left[Y_k - F(b' X_k)\right] X_k \quad (36)$$

$$\hat{l}(b) = \sum_{k \in s} \frac{1}{\pi_k} \left[y_k - F(b' x_k)\right] x_k \quad (37)$$

$$\Omega = V\left[\hat{l}(b)\right] \quad (38)$$

On remarque que la dérivée seconde de la log-vraisemblance ne dépend pas de la variable modélisée Y_k :

$$Q = \frac{\delta l}{\delta b} = \frac{\delta^2 Log(L)}{\delta b \delta b'} = -\sum_{k \in Pop} f(b' X_k) X_k X_k' = -\sum_{k \in Pop} F(b' X_k)\left[1 - F(b' X_k)\right] X_k X_k' = E\left(\frac{\delta \hat{l}}{\delta b}\right) \quad (39)$$

$\hat{b}$ étant approximativement sans biais, $(\hat{b} - b)$, où b est la vraie valeur du paramètre, est assez proche de 0 pour être linéarisé par la méthode de Taylor à l'ordre 1, en utilisant l'équation de vraisemblance :

$$\hat{l}(\hat{b}) - \hat{l}(b) \approx \frac{\delta \hat{l}}{\delta b}(\hat{b} - b) \Rightarrow \hat{l}(b) \approx -\frac{\delta \hat{l}}{\delta b}(\hat{b} - b) \text{ puisque } \hat{l}(\hat{b}) = 0$$

$$\hat{b} - b \approx -\left[\frac{\delta \hat{l}}{\delta b}\right]^{-1} \hat{l}(b)$$

Lorsque *n* augmente, $(\hat{b} - b)$ suit asymptotiquement la même distribution normale d'espérance nulle que $-Q^{-1}\left[\hat{l}(b)\right]$ [2].

D'où : $$EQM(\hat{b}) \approx \left[-\frac{\delta l}{\delta b}\right]^{-1} V\left[\hat{l}(b)\right] \left[-\frac{\delta l}{\delta b}\right]^{'-1} = Q^{-1} \Omega Q^{-1} \quad (40)$$

en remarquant que Q est une matrice symétrique, de plein rang lorsque les variables X_j ne sont pas colinéaires.

En posant : $F_k = F(b'\mathbf{x}_k)$ $\quad f_k = \dfrac{\delta F(u)}{\delta u}$, on a :

$$\Omega = V\left[\hat{l}(b)\right] = V\left[\sum_{k\in s}\frac{1}{\pi_k}\left[y_k - F(b'\mathbf{x}_k)\right]\mathbf{x}_k\right] = (\omega_j, \omega_{i,j})$$

où : $$\omega_j = \sum_{k\in \text{Pop}}\sum_{l\in \text{Pop}}\Delta_{kl}\frac{(y_k - F_k)x_{jk}}{\pi_k}\frac{(y_l - F_l)x_{jl}}{\pi_l} \qquad (41)$$

$$\omega_{ij} = \sum_{k\in \text{Pop}}\sum_{l\in \text{Pop}}\Delta_{kl}\frac{(y_k - F_k)x_{ik}}{\pi_k}\frac{(y_l - F_l)x_{jl}}{\pi_l} \qquad (42)$$

L'erreur quadratique moyenne est estimée en substituant l'estimateur $\hat{b}$ au paramètre b dans les expressions (39) à (42) et en remplaçant les totaux dans la population par leurs estimateurs Horvitz-Thomson :

$$E\hat{\text{Q}}\text{M}(\hat{b}) = \hat{Q}^{-1}\,\hat{\Omega}\,\hat{Q}^{-1} \qquad (43)$$

$$\hat{Q} = -\sum_{k\in s}\frac{1}{\pi_k}\hat{f}_k\,\mathbf{x}_k\mathbf{x}_k' = -\sum_{k\in s}\frac{1}{\pi_k}\hat{F}_k(1-\hat{F}_k)\mathbf{x}_k\mathbf{x}_k' \qquad (44)$$

avec : $\hat{F}_k = F(\hat{b}'\mathbf{x}_k)\ldots$

$$\hat{\Omega} = \hat{V}\left[\hat{l}(b)\right] = (\hat{\omega}_j, \hat{\omega}_{i,j})$$

$$\hat{\omega}_j = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{(y_k - \hat{F}_k)x_{jk}}{\pi_k}\frac{(y_l - \hat{F}_l)x_{jl}}{\pi_l}$$

$$\hat{\omega}_{ij} = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{(y_k - \hat{F}_k)x_{ik}}{\pi_k}\frac{(y_l - \hat{F}_l)x_{jl}}{\pi_l} \qquad (45)$$

3.1.3. Tests d'hypothèses sur les paramètres

On vérifie le **caractère significatif de la modalité *j*** dans le modèle en testant la nullité du paramètre b_j qui lui est affecté. La statistique de Wald, égale au carré de la statistique de Student, est utilisée pour ce test : $\dfrac{\hat{\mathbf{b}}_j^2}{\hat{V}(\hat{\mathbf{b}}_j)}$ suit un χ_1^2 (46)

Une **relation linéaire entre les paramètres** peut être formulée par : Lb=c, où L est une matrice (m,J) et c un vecteur de m constantes.

Dans un échantillon, la fonction $(\mathbf{L}\mathbf{b} - \mathbf{c})$ est estimée sans biais par $(\mathbf{L}\hat{\mathbf{b}} - \mathbf{c})$, dont la variance est égale à : $V(\mathbf{L}\hat{\mathbf{b}} - \mathbf{c}) = \mathbf{L}'V(\hat{\mathbf{b}})\mathbf{L}$. L'hypothèse H_0 est testée au moyen de la statistique :

$$W = (\mathbf{L}\hat{\mathbf{b}} - \mathbf{c})'\left[\mathbf{L}\hat{V}(\hat{\mathbf{b}})\mathbf{L}'\right]^{-1}(\mathbf{L}\hat{\mathbf{b}} - \mathbf{c}) \qquad (47)$$

qui tend vers un χ_r^2 où r est le rang de la matrice L.

3.2. Le modèle polytomique

Le modèle précédent se généralise au cas d'une variable expliquée Y prenant plus de deux modalités. On suppose que Y est une variable catégorielle à $D+1$ modalités. Pour chaque modalité *d*, on définit l'indicatrice Y_d d'appartenance à cette catégorie. Y_d est donc une variable dichotomique à valeurs 0 ou 1. On mesure, pour un individu *k*, $D+1$ variables Y_d et *J* variables X_j dont on cherche l'influence sur Y . La dernière modalité sert de référence.

Pour chaque modalité *d*, on pose :

$$p_{dk} = \text{Prob}\{Y_{dk} = 1/\mathsf{X}_k\} = \text{Prob}\{\theta_d'\mathsf{X}_k + U_k > 0\} = F(\theta_d'\mathsf{X}_k) \quad (48)$$

$$p_{(D+1)k} = F(\theta_{(D+1)}'\mathsf{X}_k) = 1 - \sum_{d=1}^{D} F(\theta_d'\mathsf{X}_k)$$

où F est la fonction de répartition multinomiale du résidu du modèle. On estime un vecteur de paramètres θ_d de dimension *J+1* incluant un terme constant pour chaque modalité *d* de Y.

Le modèle logit multinomial a pour expression : $F(\theta_d'\mathsf{X}_k) = \dfrac{\exp(\theta_d'\mathsf{X}_k)}{1+\sum_{d=1}^{D}\exp(\theta_d'\mathsf{X}_k)}$ (49).

La log-vraisemblance devient (dans la population) :

$$Log(L) = \sum_{k\in \text{Pop}}\left[\sum_{d=1}^{D}\left(Y_{dk}\,Log\left[F(\theta_d'\mathsf{X}_k)\right]\right) + Y_{(D+1)k}\,Log\left(1-\sum_{d=1}^{D}F(\theta_d'\mathsf{X}_k)\right)\right]$$

qu'on peut résumer par :

$$Log(L) = \sum_{k\in \text{Pop}}\left(\mathsf{Y}_k'Log[\mathsf{p}_k] + Y_{(D+1)k}\,Log\left[1-\sum_{d=1}^{D}F(\theta_d'\mathsf{X}_k)\right]\right)$$

où p_k est le vecteur des *D* probabilités définies en (48) et Y_k le vecteur des *D* premières indicatrices Y_d pour l'individu *k*. Elle est estimée dans l'échantillon par :

$$\hat{Log}(L) = \sum_{k\in \mathsf{s}}\frac{1}{\pi_k}\left(\mathsf{y}_k'Log[\mathsf{p}_k] + y_{(D+1)k}\,Log\left[1-\sum_{d=1}^{D}F(\theta_d'\mathsf{X}_k)\right]\right)$$

Les estimateurs des paramètres θ_d sont les solutions des équations de vraisemblance :

$$\frac{\delta\,\hat{Log}(L)}{\delta\,\theta_d} = 0$$

3.3. La procédure SURVEYLOGISTIC

A condition de spécifier les poids de sondage en variable de pondération, la procédure SURVEYLOGISTIC permet de prendre en compte le plan de sondage dans l'estimation des paramètres et de leur précision. Elle propose l'ajustement des modèles dichotomique, polytomique ordonné et multinomial.

3.3.1. L'estimation des paramètres

La procédure SURVEYLOGISTIC propose sur option **quatre fonctions de lien**. Les trois premières utilisent la fonction de répartition logistique spécifiée en (32), mais se distinguent par leur mode de traitement du modèle polytomique.

- la fonction logit généralisée (*generalized logit function*) est celle spécifiée en (49), qui permet d'ajuster un modèle dichotomique ou un modèle multinomial
- la fonction logit cumulée (*cumulative logit function*), option par défaut de la procédure, produit les mêmes probabilités qu'en (32) dans le cas du modèle dichotomique, mais a pour expression : $F\left(\theta_d^{'} X_k\right) = \sum_{r=1}^{d} p_{rk}$ où $p_{rk} = E\left(y_{rk} / x_k\right) = \text{Prob}\left\{y_{rk} = 1 / x_k\right\}$ dans le cas d'un modèle polytomique.
- la fonction log-log s'écrit : $Log\left(-Log\left[1 - F\left(\theta_d^{'} X_k\right)\right]\right) = \alpha_d + \beta^{'} x_k$ avec $\alpha_1 < \alpha_2 < ... < \alpha_D$
- le modèle probit utilise la fonction de répartition de la loi normale.

Seule la première fonction de lien permet d'ajuster un modèle multinomial, impliquant des coefficients distincts selon les modalités de la variable Y pour un même paramètre explicatif X_j. Les autres fonctions estiment des modèles polytomiques ordonnés.

Par défaut, SAS résout les équations de vraisemblance par **l'algorithme** itératif de Fisher (« Fisher scoring »). De façon optionnelle, la procédure peut aussi utiliser la méthode Newton-Raphson. Celle-ci est utilisée obligatoirement avec la fonction de lien du modèle logit généralisé. Les deux méthodes coïncident lorsqu'on utilise la fonction logit en fonction de lien avec un modèle dichotomique.

3.3.2. La précision de l'estimateur

Lorsque le paramètre $\hat{\beta}$ est calculé par la méthode de Fisher, le logiciel SAS estime sa variance au moyen de la formule générale ci-dessous, adaptée au modèle polytomique.

Soient : $F_{kd} = F\left(\theta_d^{'} x_k\right)$

$D_k(D,J) = \left[... \frac{\delta F_{kd}}{\delta \theta_{d_j}} ... \right]$ est la matrice des dérivées partielles par rapport aux paramètres à estimer

$P_k = Diag\left(\hat{p}_k\right)$ est la matrice diagonale dont les éléments non nuls sont les coordonnées du vecteur $\hat{p}_k$ défini en section 3.2 et estimé au point $\hat{\theta}$.

Variance estimée des coefficients de régression :

$$\boxed{\hat{V}\left(\hat{\theta}\right) = \hat{Q}^{-1} \hat{G} \hat{Q}^{-1}}$$

$$\hat{Q} = \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{k=1}^{n_{hi}} \frac{1}{\pi_k} \hat{D}_{hik} \left[P_{hik} - \hat{p}_{hik} \hat{p}_{hik}^{'}\right]^{-1} \hat{D}_{hik}^{'} \qquad (50)$$

$$\hat{\mathrm{G}} = \frac{n-1}{n-p}\sum_{h=1}^{H}\frac{m_h\left(1-f_h\right)}{m_h-1}\sum_{i=1}^{m_h}\left(e_{hi.}-\bar{e}_{h..}\right)\left(e_{hi.}-\bar{e}_{h..}\right)' \qquad (51)$$

$$e_{hi.} = \sum_{k=1}^{n_{hi}}\frac{1}{\pi_k}\hat{\mathrm{D}}_{hik}\left[\mathrm{P}_{hik}-\hat{\mathrm{p}}_{hik}\hat{\mathrm{p}}_{hik}'\right]^{-1}\left(y_{hik}-\hat{\mathrm{p}}_{hik}\right) \qquad (52)$$

$$\bar{e}_{h..} = \frac{1}{m_h}\sum_{i=1}^{m_h} e_{hi.} \qquad (53)$$

h est l'indice de la strate, i l'indice de l'unité primaire et k l'indice de l'individu. La strate h de l'échantillon comprend m_h unités primaires et l'unité primaire i contient n_{hi} individus.

f_h est le taux de sondage dans la strate h. Lorsque le sondage est à plusieurs degrés, c'est le taux de sondage au premier degré de tirage.

Lorsque les équations de vraisemblance sont résolues par la méthode de Newton-Raphson, la matrice Q est remplacée par l'espérance de l'opposé du Hessien, matrice des dérivées secondes de la log-vraisemblance : $E\left(-\hat{\mathrm{H}}\right) = E\left[-\frac{\delta^2 L\hat{o}g(L)}{\delta\hat{\theta}\,\delta\hat{\theta}'}\right]$.

On retrouve, dans les expressions ci-dessus, les formules (43) à (45) lorsque le modèle est dichotomique. Le calcul de la matrice G repose sur les approximations habituelles du logiciel SAS en cas de sondage complexe, déjà décrites dans la section 1.2.2.

Dans l'expression de $\hat{\mathrm{G}}$, le facteur $(n\text{-}1)/(n-p)$, où p est le nombre de paramètres du modèle, est destiné à corriger le biais d'estimation de la variance dans les petits échantillons par le nombre de degrés de liberté. C'est la méthode utilisée par défaut.

La procédure comprend également une option d'ajustement par la méthode proposée par J.G.Morel [7], qui conduit à :

$$\hat{V}\left(\hat{\theta}\right) = \hat{Q}^{-1}\hat{G}\hat{Q}^{-1} + k\lambda\,\hat{Q}^{-1} \qquad \text{avec :} \quad k = \max\left(\delta, \frac{1}{p}tr\left(\hat{Q}^{-1}\hat{G}\right)\right) \quad \lambda = \min\left(\phi, \frac{p}{m-p}\right)$$

δ et ϕ sont paramétrables. Par défaut, ils valent respectivement 1 et 0,5.

3.3.3. Test d'hypothèses sur les paramètres

Lorsque la variable d'intérêt Y a plus de deux modalités, la procédure teste **l'égalité des coefficients θ_{dj} affectés au même facteur explicatif** X_j. Selon la fonction de lien utilisée, ce test est appelé « test d'égalité des pentes » (*score test for the equal slopes assumption*) ou « test de proportionnalité des odd-ratios » (*score test for the proportional odds assumption*). $\hat{\theta}_0$ étant le vecteur des paramètres estimés sous l'hypothèse d'égalité des pentes, et $g = F^{-1}$ la fonction inverse de la fonction de répartition, ce test utilise la statistique :

$$g'\left(\hat{\theta}_0\right)\left[-\hat{H}\left(\hat{\theta}_0\right)\right]^{-1} g\left(\hat{\theta}_0\right)$$

qui suit un $\chi^2_{J(D-1)}$ où J est le nombre de facteurs explicatifs X_j du modèle et $(D+1)$ le nombre de modalités de la variable expliquée.

Une **hypothèse linéaire générale** sur les paramètres de la forme H_0 : Lb=c, où L est une matrice (*m*,*J*) et c un vecteur de *m* constantes est testée comme indiqué à la section 3.1.3.

3.4. Un exemple de modélisation avec SURVEYLOGISTIC

C'est toujours l'enquête santé de 2001 qui sert de cadre de référence aux tests suivants. On modélise la propension à fumer selon divers critères démographiques, sociaux, sanitaires. La variable expliquée (*FUMEUR*), dichotomique, prend les valeurs 1 pour un individu fumeur et 0 pour un individu non fumeur.

On a retenu les facteurs suivants :

- le sexe
- l'âge en 3 modalités :
  - moins de 40 ans
  - 40 à 64 ans
  - 65 ans et plus
- l'exercice d'une activité professionnelle
  - inactif
  - chômeur
  - actif exerçant un emploi
- la sensation de stress :
  - stress au travail
  - stress dans la vie personnelle
  - absence de stress
- l'état de santé général :
  - bon
  - moyen
  - médiocre ou mauvais.

Un exemple de programme avec la procédure SURVEYLOGISTIC :

```
PROC SURVEYLOGISTIC DATA=echant RATE=plan;
     STRATA sante ;
     CLASS  sexe age1 stress activite sante;
     MODEL fumeur (EVENT='1') = sexe age1 stress activite sante;
     WEIGHT poids;
     FORMAT age1 $trag3x. sexe $sexe. stress $str. activite $activ.;
     TITLE  "Sondage aléatoire simple";
     ODS OUTPUT TYPE3=effet PARAMETERESTIMATES=param;
 RUN;
```

Dans cet exemple, l'échantillon est à un degré stratifié selon la variable *sante*, et a été sélectionné par sondage aléatoire simple dans les strates. Les taux de sondage par strate, contenus dans la table *plan*, sont déclarés dans une option RATE=*plan* pour que les variances de strate soient multipliées par $(1-f_h)$. Le critère de stratification est déclaré dans l'instruction STRATA et le poids de sondage dans l'instruction WEIGHT. L'instruction ODS OUTPUT stocke dans une table *effet* les résultats des tests de type III de significativité des facteurs, et dans une table *param* la valeur des coefficients de régression avec les statistiques qui leur sont associées.

Le modèle a pour catégorie de référence les hommes de 40 à 64 ans, exerçant un emploi, ne ressentant pas de stress et percevant leur état de santé comme moyen. Le modèle tente d'établir des corrélations mais ne préjuge pas du sens de la causalité entre les variables exogènes et la variable expliquée. Les deux derniers facteurs en particulier – stress et perception de son état de santé – peuvent être interprétés comme la conséquence plutôt que la cause de la propension à fumer.

Les paramètres ont été estimés successivement dans la population de référence avec la procédure LOGISTIC sans pondération, puis dans chaque échantillon par les procédures LOGISTIC et SURVEYLOGISTIC, en pondérant les observations par leur poids de sondage dans les deux cas. Les résultats obtenus dans chaque série de 100 échantillons sont résumés par les médianes des X-value associées à chaque effet et des X-value associées à chaque paramètre.

3.4.1. Sondage aléatoire simple

Paramètres	**Population de référence**	**LOGISTIC pondéré**	**SURVEYLOGISTIC**
Test des effets : χ^2-Wald			
• sexe	67,90	6,29	6,78
• âge	219,27	19,08	20,40
• stress	12,14	1,90	2,08
• santé	12,19	2,31	2,63
• activité	32,02	3,97	4,34
Ecarts-types des coefficients de régression :			
• Intercept	0,0698	0,2332	0,2198
• Femmes	0,0342	0,1127	0,1085
• Moins de 40 ans	0,0598	0,1966	0,1908
• 65 ans et plus	0,0913	0,2992	0,2909
• stress au travail	0,0526	0,1727	0,1657
• stress dans la vie	0,0604	0,2007	0,1927
• bonne santé	0,0623	0,2086	0,1986
• mauvaise santé	0,0952	0,3238	0,3049
• inactif	0,0656	0,2187	0,2091
• chômeur	0,0944	0,3169	0,3023
χ^2-Wald des coefficients de régression :			
• Intercept	268,45	24,99	28,30
• Femmes	67,90	6,29	6,78
• Moins de 40 ans	217,83	18,60	19,52
• 65 ans et plus	155,99	14,05	14,92
• stress au travail	0,65	0,57	0,64
• stress dans la vie	5,04	0,52	0,53
• bonne santé	4,15	0,35	0,38
• mauvaise santé	0,21	0,29	0,39
• inactif	32,02	3,02	3,39
• chômeur	16,71	1,80	1,82

3.4.2. Sondage stratifié avec sondage aléatoire simple dans les strates

Paramètres	**Population de référence**	**LOGISTIC pondéré**	**SURVEYLOGISTIC**
Test des effets : χ^2-Wald			
• sexe	67,90	6,89	6,97
• âge	219,27	21,26	20,68
• stress	12,14	2,16	2,20
• santé	12,19	2,62	2,66
• activité	32,02	3,93	4,16
Ecarts-types des coefficients de régression :			
• Intercept	0,0698	0,2349	0,2284
• Femmes	0,0342	0,1125	0,1112
• Moins de 40 ans	0,0598	0,1970	0,1973
• 65 ans et plus	0,0913	0,3026	0,3033
• stress au travail	0,0526	0,1721	0,1692
• stress dans la vie	0,0604	0,1987	0,1954
• bonne santé	0,0623	0,2064	0,2034
• mauvaise santé	0,0952	0,3138	0,3099
• inactif	0,0656	0,2189	0,2150
• chômeur	0,0944	0,3169	0,3080
χ^2-Wald des coefficients de régression :			
• Intercept	268,45	25,50	27,32
• Femmes	67,90	6,89	6,97
• Moins de 40 ans	217,83	20,22	20,33
• 65 ans et plus	155,99	15,08	14,50
• stress au travail	0,65	0,28	0,30
• stress dans la vie	5,04	1,06	1,10
• bonne santé	4,15	0,57	0,61
• mauvaise santé	0,21	0,48	0,49
• inactif	32,02	2,81	2,94
• chômeur	16,71	1,18	1,30

3.4.3. Sondage à deux degrés

Le plan de sondage est différent de la précédente simulation. Il faut en effet avoir, dans l'échantillon, un nombre d'unités primaires supérieur d'au moins deux unités au nombre de degrés de liberté du modèle pour que la variance des paramètres puisse être estimée. On a donc ici sélectionné 11 régions au premier degré, puis 46 individus par région au second degré.

Paramètres	**Population de référence**	**LOGISTIC pondéré**	**SURVEYLOGISTIC**
Test des effets : X^2-Wald			
• sexe	67,90	6,91	16,39
• âge	219,27	20,46	45,68
• stress	12,14	2,05	5,62
• santé	12,19	2,33	6,64
• activité	32,02	4,15	12,39
Ecarts-types des coefficients de régression :			
• Intercept	0,0698	0,2304	0,1504
• Femmes	0,0342	0,1117	0,0742
• Moins de 40 ans	0,0598	0,1964	0,1436
• 65 ans et plus	0,0913	0,2982	0,2105
• stress au travail	0,0526	0,1735	0,1065
• stress dans la vie	0,0604	0,1977	0,1372
• bonne santé	0,0623	0,2022	0,1333
• mauvaise santé	0,0952	0,3051	0,2202
• inactif	0,0656	0,2147	0,1369
• chômeur	0,0944	0,3052	0,2068
X^2-Wald des coefficients de régression :			
• Intercept	268,45	26,21	66,25
• Femmes	67,90	6,91	16,38
• Moins de 40 ans	217,83	19,61	38,89
• 65 ans et plus	155,99	14,63	32,51
• stress au travail	0,65	0,52	1,36
• stress dans la vie	5,04	0,63	1,24
• bonne santé	4,15	0,83	1,56
• mauvaise santé	0,21	0,55	1,09
• inactif	32,02	2,89	6,73
• chômeur	16,71	2,14	5,62

Bibliographie

[1] P.D. Alllison , “Logistic Regression using the SAS system”, *SAS Institute*, 1999.

[2] D.A. Binder, “On the variances of asymptotically normal estimators from complex surveys”, *International Statistical Review*, vol.51 1983.

[3] P. Druilhet, “Analyse de la variance“, *polycopié ENSAI*, 2000.

[4] C. Gouriéroux, “Econométrie des variables qualitatives“, 1984.

[5] R. Lehtonen, E.J. Pahkinen, “Practical methods for design and analysis of complex surveys”, 1994.

[6] M.Marpsat et alii, “L’économétrie et l’étude des comportements“, *INSEE, document de travail n°0001*, 2000.

[7] J.G. Morel, “Régression logistique selon des plans de sondage complexes”, *Techniques d’enquête*, vol.15, décembre 1989.

[8] P. Peretti-Watel, “Régression sur variables catégorielles“, *polycopié ENSAI*, 2000.

[9] G. Roberts, J.N.K. Rao, S. Kumar, “Logistic regression analysis of sample survey data”, *Biometrika*, mars1987.

[10] E. Särndal, “Model assisted survey sampling”, 1992.

[11] N.H. Timm et T. A. Mieczkowski, “General linear models, theory and applications using SAS software”, *SAS Institute*, 1997.

[12] "SAS-Stat User’s guide", *SAS Institute.*