La régression sur échantillon avec le logiciel SAS

1) Régression linéaire

- Population $P = \{i_1, \ldots i_N\}$
- Echantillon $s = \{i_1, \ldots i_n\}$

  probabilité d'inclusion : $\text{Prob}\{i_k \in s\} = \pi_k$
- On suppose que dans la population :

$$Y_k = \sum_{j=1}^{p} b_j X_{jk} + U_k = b' X_k + U_k \quad \Leftrightarrow Y = X \quad b + u$$

$$(N,1) \quad (N,p)(p,1) \quad (N,1)$$

et que cette relation est vérifiée dans l'échantillon

- Le problème se ramène à estimer, à partir de l'échantillon, la valeur des coefficients de la régression dans la population
- L'estimateur des moindres carrés dans l'univers :

$$\tilde{b}' = (X'X)^{-1}X'Y = T^{-1}t$$

$$\alpha_{ij}(T) = \sum_{k \in U} X_{ik} X_{jk}$$

$$\beta_j(t) = \sum_{k \in U} X_{jk} Y_k$$

- L'estimateur Horvitz-Thomson des éléments de T et t :

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

- L’estimateur des coefficients de régression :

$$\hat{b} = \hat{T}^{-1}\hat{t} = \left(X_s' W X_s\right)^{-1} X_s' W Y_s$$

- Espérance et variance de l’estimateur

$$\hat{b} - \tilde{b} \approx \frac{\delta \tilde{b}}{\delta t}\left(\hat{t} - t\right) + \frac{\delta \tilde{b}}{\delta T}\left(\hat{T} - T\right) = T^{-1}\left(\hat{t} - \hat{T}\tilde{b}\right) = T^{-1}\left(X_s' W \tilde{U}_s\right)$$

- Eléments de $\left(X_s' W \tilde{U}_s\right)$: $Z_i = \sum_{k \in s} \frac{x_{ik}\tilde{u}_k}{\pi_k} = \sum_{k \in s} \frac{z_{ik}}{\pi_k}$

$$E\left(\hat{b}\right) \approx \tilde{b}$$

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

$$V_{ij} = \sum_{k \in U}\sum_{l \in U} \Delta_{kl} \frac{X_{ik}\tilde{U}_k}{\pi_k}\frac{X_{jl}\tilde{U}_l}{\pi_l}$$

Variance estimée de $\hat{b}$

$$E\hat{Q}M\left(\hat{b}\right) = \left(X_s'WX_s\right)^{-1}\left[\hat{V}\right]\left(X_s'WX_s\right)^{-1}$$

$$\hat{V} = \left(V_{ij}\right) = \left(\sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{x_{ik}\hat{u}_k}{\pi_k} \frac{x_{jl}\hat{u}_l}{\pi_l} \right)$$

La procédure SURVEYREG

```
PROC SURVEYREG DATA=echantillon RATE=plan;
   CLUSTER ident ;
   STRATA region  type ;
   WEIGHT poids ;
   CLASS pcs;
   MODEL consom=revenu age pcs   / SOLUTION;
RUN;
```

La procédure SURVEYREG

$$\hat{\mathbf{b}}_{SAS} = \left(\mathbf{X}_s'\mathbf{W}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathbf{W}\mathbf{Y}_s$$

avec : $\mathbf{W} = Diag\left(w_k\right)$

$$\hat{V}_{SAS}\left(\hat{\mathbf{b}}\right) = \left(\mathbf{X}_s'\mathbf{W}\mathbf{X}_s\right)^{-1}\mathbf{G}\left(\mathbf{X}_s'\mathbf{W}\mathbf{X}_s\right)^{-1}$$

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

$$\mathbf{e}_{hik} = w_{hik}\hat{u}_{hik}\mathbf{X}_{hik}$$

$$\mathbf{e}_{hi.} = \sum_{k=1}^{n_{hi}}\mathbf{e}_{hik}$$

$$\bar{\mathbf{e}}_{h..} = \frac{1}{m_h}\sum_{i=1}^{m_h}\mathbf{e}_{hik}$$

Sondage aléatoire simple

$$G = N^2 \frac{(1-f)}{n} \frac{(n-1)}{n-p} \left[s^2 \right]$$

$$\left[s^2 \right] = \left(\ldots s_j^2 \ldots \right) = \left(\ldots \frac{1}{n-1} \sum_{k \in s} \left(z_{jk} - \bar{z}_j \right)^2 \ldots \right)$$

$$\left(\ldots s_{jq} \ldots \right) = \left(\ldots \frac{1}{n-1} \sum_{k \in s} \left(z_{jk} - \bar{z}_j \right) \left(z_{qk} - \bar{z}_q \right) \ldots \right)$$

$$z_{jk} = \hat{u}_k x_{jk}$$

Les autres plans de sondage

Les hypothèses de calcul

- **sondage à probabilités proportionnelles à la taille** : avec remise
- **sondage en grappes** : formules habituelles
- **sondage à deux degrés** : premier degré avec remise ↯ estimation de la variance inter-UP seulement
- **sondage stratifié** : la variance est la somme des variances de strate

Test de nullité d'un paramètre

$\hat{T}=\frac{\hat{b}_j}{\hat{\sigma}_{\hat{b}_j}}$ suit approximativement une loi

de Student à (m-H) degrés de liberté où

- m=nombre d'unités primaires de l'échantillon
- H=nombre de strates

- (m=n dans un sondage à 1 degré et H=1 dans un sondage non stratifié)

2) Analyse de la variance

Le modèle : $$Y_k = \sum_{i=1}^{p} \alpha_i X_{ik} + \sum_{j=1}^{q} \beta_j Y_{jk} + \sum_{ij=1}^{pq} \gamma_{ij} T_{ijk} + U_k = \mathsf{A}\theta + \mathsf{u}$$

Les variables X_i ,Y_j sont des indicatrices d'appartenance à une catégorie

Test d'une hypothèse linéaire sur les paramètres H0 : $\mathsf{c}'\theta = 0$

Dans la population :

$$F = \frac{1}{p_0}\left(\mathsf{c}'\tilde{\theta}\right)'\left[\mathsf{c}'\tilde{V}\left(\tilde{\theta}\right)\mathsf{c}\right]^{-1}\left(\mathsf{c}'\tilde{\theta}\right) \quad \text{suit un} \quad \mathcal{F}_{(p_0, N-p)}$$

où p_o=nombre de paramètres sous H0

Avec SURVEYREG

SURVEYREG réalise les tests de type III (modèle complet contre modèle privé de l'un des facteurs) et teste les fonctions estimantes.

Dans l'échantillon :

$$\hat{F} = \frac{1}{p_0}\left(\mathbf{c}'\hat{\boldsymbol{\theta}}\right)'\left[\mathbf{c}'\hat{V}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{c}\right]^{-1}\left(\mathbf{c}'\hat{\boldsymbol{\theta}}\right)$$ suit approximativement

un $\mathcal{F}_{(p_0, m-H)}$

où p_o= nombre de paramètres sous H0

m=nombre d'unités primaires dans l'échantillon

H=nombre de strates

Contraintes identifiantes

- sans interaction : annulation de la dernière modalité de chaque effet à partir du 2ème
- avec interaction : annulation de la dernière modalité des effets à partir du 2ème et des (p_1+…+p_q) dernières modalités croisées

Un exemple : l'enquête PCV 2001, volet santé

- *Variable expliquée* :
  - nombre de consultations auprès d'un généraliste dans l'année
- *Facteurs explicatifs* :
  - l'âge : moins de 50 ans, 50-79 ans, 80 ans et +
  - le sexe
  - le niveau de couverture sociale : pas de mutuelle, mutuelle, CMU
- *Simulations* : - sondage aléatoire simple
  - sondage stratifié avec PESR dans les strates

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

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

3) La régression logistique

- Y est une variable dichotomique à expliquer
- Y^* une variable latente
- $X_1,...X_p$ p variables exogènes observées
- $b_1,...b_p$ des coefficients

Le modèle : $p_k = \text{Prob}\{Y_k = 1\} = \text{Prob}\{Y_k^* \geq 0\} = \text{Prob}\{b'X_k + u_k \geq 0\} = F(b'X_k)$

Dans la population, les $\tilde{b}_j$ sont les solutions des équations :

$$\frac{\delta Log(L)}{\delta b} = \sum_{k=1}^{N} \frac{Y_k - F(b'X_k)}{F(b'X_k)(1-F(b'X_k))} f(b'X_k) X_k = 0 \quad \text{où} : f = \frac{\delta F}{\delta u}$$

Variance des coefficients :

$$V(\tilde{b}) = \left[\sum_{k=1}^{N} \frac{[f(b'X_k)]^2}{F(b'X_k)(1-F(b'X_k))} X_k X'_k \right]^{-1}$$

Dans l'échantillon, on résout les équations :

$$\frac{\delta Log\left(\hat{L}\right)}{\delta\,\hat{b}} = \sum_{k\in s}\frac{1}{\pi_k}\frac{y_k - F\left(\hat{b}'x_k\right)}{F\left(\hat{b}'x_k\right)\left(1-F\left(\hat{b}'x_k\right)\right)} f\left(\hat{b}'x_k\right)x_k = 0$$

Avec une fonction logit, l'erreur quadratique moyenne est estimée par :

$$E\hat{Q}M\left(\hat{b}\right) = \left(X_s'\hat{\Delta}X_s\right)^{-1}\left(X_s'\hat{V}X_s\right)\left(X_s'\hat{\Delta}X_s\right)^{-1}$$

$$\hat{\Delta} = Diag\left[\frac{F\left(b'x_k\right)\left(1-F\left(b'x_k\right)\right)}{\pi_k}\right]$$

$$\hat{V} = \hat{V}\left(y_k - F\left(b'x_k\right)\right)$$

C'est cette formule qu'applique SURVEYLOGISTIC avec, pour $\hat{V}$, les mêmes approximations selon le plan de sondage.

La procédure SURVEYLOGISTIC

```
PROC SURVEYLOGISTIC DATA=echantillon RATE=plan;
    CLUSTER ident;
    STRATA region type;
    WEIGHT poids;
    CLASS sexe age activite stress ;
    MODEL fumeur (EVENT='1') = sexe age activite stress ;
   RUN;
```

Un exemple : l'enquête PCV 2001, volet santé

- *Variable expliquée* :
  - propension à fumer quotidiennement
- *Facteurs explicatifs* :
  - l'âge : moins de 40 ans, 40-64 ans, 65 ans et +
  - le sexe
  - le type d'activité : en emploi, chômeur, inactif
  - le stress éprouvé : stress au travail, stress dans la vie personnelle, pas de stress
  - l'état de santé ressenti : bon, moyen, mauvais
- *Catégorie de référence* : hommes de 40 à 64 ans, exerçant un emploi, ne ressentant pas de stress et percevant leur état de santé comme moyen
- *Simulation* de sondage stratifié avec PESR 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