

# **La régression sur échantillon avec le logiciel SAS**

# 1) Régression linéaire

- Population  $P = \{i_1, \dots, i_N\}$
- Echantillon  $S = \{i_1, \dots, i_n\}$

probabilité d'inclusion :  $\text{Prob}\{i_k \in S\} = \pi_k$

- On suppose que dans la population :

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

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

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 :

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

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

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

- L'estimateur Horvitz-Thomson des éléments de  $\mathbf{T}$  et  $\mathbf{t}$  :

$$\hat{\alpha}_{ij} = \sum_{k \in S} \frac{x_{ik} x_{jk}}{\pi_k}$$

$$\hat{\beta}_j = \sum_{k \in S} \frac{x_{jk} y_k}{\pi_k}$$

- L'estimateur des coefficients de régression :

$$\hat{\mathbf{b}} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{t}} = (\mathbf{X}'_s \mathbf{W} \mathbf{X}_s)^{-1} \mathbf{X}'_s \mathbf{W} \mathbf{Y}_s$$

- Espérance et variance de l'estimateur

$$\hat{\mathbf{b}} - \tilde{\mathbf{b}} \approx \frac{\partial \hat{\mathbf{b}}}{\partial \hat{\mathbf{t}}} (\hat{\mathbf{t}} - \mathbf{t}) + \frac{\partial \hat{\mathbf{b}}}{\partial \hat{\mathbf{T}}} (\hat{\mathbf{T}} - \mathbf{T}) = \mathbf{T}^{-1} (\hat{\mathbf{t}} - \hat{\mathbf{T}} \tilde{\mathbf{b}}) = \mathbf{T}^{-1} (\mathbf{X}'_s \mathbf{W} \tilde{\mathbf{U}}_s)$$

- Éléments de  $(\mathbf{X}'_s \mathbf{W} \tilde{\mathbf{U}}_s)$  :  $Z_i = \sum_{k \in S} \frac{x_{ik} \tilde{u}_k}{\pi_k} = \sum_{k \in S} \frac{Z_{ik}}{\pi_k}$

$$E(\hat{\mathbf{b}}) \approx \tilde{\mathbf{b}}$$

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

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

Variance estimée de  $\hat{b}$

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

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

# 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

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

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

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

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

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

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

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

# Sondage aléatoire simple

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

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

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

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



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

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

Test d'une hypothèse linéaire sur les paramètres  $H_0 : c'\theta = 0$

**Dans la population :**

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

où  $p_0$  = nombre de paramètres sous  $H_0$

# 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 :

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

un  $F_{(p_0, m-H)}$

où  $p_0$  = 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  $(p_1 + \dots + p_q)$  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

<b>Paramètres</b>	<b>Population de référence</b>	<b>GLM pondéré</b>	<b>SURVEYREG</b>
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
Ecart-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

<b>Paramètres</b>	<b>Population de référence</b>	<b>GLM pondéré</b>	<b>SURVEYREG</b>
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
- $X_1, \dots, X_p$   $p$  variables exogènes observées
- $b_1, \dots, b_p$  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 :

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

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 :

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

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

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

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

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

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

<b>Paramètres</b>	<b>Population de référence</b>	<b>LOGISTIC pondéré</b>	<b>SURVEYLOGISTIC</b>
Test des effets : $\chi^2$ -Wald <ul style="list-style-type: none"> <li>• sexe</li> <li>• âge</li> <li>• stress</li> <li>• santé</li> <li>• activité</li> </ul>	67,90 219,27 12,14 12,19 32,02	6,89 21,26 2,16 2,62 3,93	6,97 20,68 2,20 2,66 4,16
Ecarts-types des coefficients de régression : <ul style="list-style-type: none"> <li>• Intercept</li> <li>• Femmes</li> <li>• Moins de 40 ans</li> <li>• 65 ans et plus</li> <li>• stress au travail</li> <li>• stress dans la vie</li> <li>• bonne santé</li> <li>• mauvaise santé</li> <li>• inactif</li> <li>• chômeur</li> </ul>	0,0698 0,0342 0,0598 0,0913 0,0526 0,0604 0,0623 0,0952 0,0656 0,0944	0,2349 0,1125 0,1970 0,3026 0,1721 0,1987 0,2064 0,3138 0,2189 0,3169	0,2284 0,1112 0,1973 0,3033 0,1692 0,1954 0,2034 0,3099 0,2150 0,3080
$\chi^2$ -Wald des coefficients de régression : <ul style="list-style-type: none"> <li>• Intercept</li> <li>• Femmes</li> <li>• Moins de 40 ans</li> <li>• 65 ans et plus</li> <li>• stress au travail</li> <li>• stress dans la vie</li> <li>• bonne santé</li> <li>• mauvaise santé</li> <li>• inactif</li> <li>• chômeur</li> </ul>	268,45 67,90 217,83 155,99 0,65 5,04 4,15 0,21 32,02 16,71	25,50 6,89 20,22 15,08 0,28 1,06 0,57 0,48 2,81 1,18	27,32 6,97 20,33 14,50 0,30 1,10 0,61 0,49 2,94 1,30