STAT0162-1 Analyse statistique de données qualitatives et quantitatives en sciences sociales

Transparents

Philippe Lambert

http : //www.statsoc.ulg.ac.be/quali.html

Institut des Sciences Humaines et Sociales
Université de Liège

Analyse de tables de contingence: le modèle log-linéaire

Contexte

• Le modèle de régression log-linéaire est souvent utilisé pour modéliser les associations entre variables catégorielles.

• Le point de départ est habituellement une table de contingence.

Ex European Social Survey (année 2002 pour la Belgique) - Intéressons-nous à la question suivante:

Avez-vous déjà été sans emploi et à la recherche d'un travail pendant une période de plus de trois mois ? 1: oui 2: non 7: refus 8: sait pas 9: sans réponse.

Contexte (2)

• Sur base du genre et de l'âge des personnes questionnées, on peut construire la table des fréquences (appelée *table de contingence*) suivante:

		Age			
Sexe	Chômage	30-39	40-49	50-65	65+
Homme	Oui	59	52	31	14
	Non	113	106	146	110
Femme	Oui	72	76	47	11
	Non	71	91	158	135

L'information sur les 3 critères est disponible pour 1292 personnes.

[Données extraites de `ESS2002BEbis.sta` et reprises dans `chomage.sta`].

• De nombreuses questions peuvent être posées au départ d'une telle table à propos des liens éventuels entre les (ou un sous-ensemble des) variables impliquées.

Réponse binaire

• Lorsqu'une variable réponse **binaire** est désignée, la régression logistique permet d'évaluer la nature des liens l'unissant aux autres variables.

Ex Comment évolue la probabilité d'avoir été sans emploi et à la recherche d'un travail pendant une période de plus de trois mois avec le `Sexe`?

• Les données peuvent être réorganisées sous la forme d'une table 2x2 [`chomage2.sta`]:

	Chômage		
Sexe	Non	Oui	Total
Femme	455 (69%)	206 (31%)	661 (100%)
Homme	475 (75%)	156 (25%)	631 (100%)

• Le modèle de régression logistique décrit comment le logarithme de la cote du chômage change avec `Sexe`:

$$\log \frac{\pi_i}{1-\pi_i} = \theta_i = \mu + \alpha_i \ ; \ (\alpha_1 = 0)$$

Chomage - Paramètres estimés (chomage2.sta)
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	-0,792421	0,083977	89,04087	0,000000
Sexe	Homme	2	-0,321038	0,124771	6,62045	0,010081
Sexe	Femme	3	0,000000			
Echelle			1,000000	0,000000		

• La catégorie 'Femme' a été désignée comme référence ($i = 1$):

$$\hat{\mu} = -0.79 = \log(\widehat{\text{Cote}}_F) \; ; \; \hat{\alpha}_2 = -0.32 = \log(\widehat{\text{RC}})$$

• La P-valeur associée à $\hat{\alpha}_2$, $0.01 < 5\%$, indique que 0 ne fait pas partie de l'intervalle de crédibilité 95%. L'égalité des cotes du chômage chez les hommes et les femmes ($\text{RC} = 1 \Rightarrow \log(RC) = 0$) n'est donc pas une hypothèse plausible: il existe un **lien significatif** entre Sexe et le risque d'être chômeur (au sens défini par le questionnaire).

• Pour rappel, cette prise de décision, reposant sur une approximation normale de la distribution a posteriori de α_2, doit idéalement être validée par un test chi-carré.

• La cote du chômage chez les hommes est significativement plus faible que chez les femmes et estimée égale à 73% ($= \exp(-0.32) = 0.73$) de cette dernière.

• La tendance observée dans la table de départ (cf. pourcentages) n'est donc pas imputable au hasard.

Table de contingence: notations

• Le point de départ est la table des fréquences [`chomage2.sta`]:

	Chômage		
Sexe	$(j=1)$ Non	$(j=2)$ Oui	Total
$(i=1)$ Femme	n_{11} =455	n_{12} =206	n_{1+} =661
$(i=2)$ Homme	n_{21} =475	n_{22} =156	n_{2+} =631
	n_{+1} =930	n_{+2} =362	n_{++} =1292

• L'analyse peut également se faire conditionnellement au nombre de personnes questionnées avec une information complète pour les critères d'intérêt, $n_{++} = 1292$.

Notations:

→ π_{ij}: proportion, dans la population, de personnes dans la cellule (i, j).

→ $\mu_{ij} = n_{++}\pi_{ij}$: nombre de personnes attendues (en moyenne. . . si on répète de telles enquêtes) dans la cellule (i, j).

	$j=1$	$j=2$	Total
$i=1$	π_{11}	π_{12}	π_{1+}
$i=2$	π_{21}	π_{22}	π_{2+}
	π_{+1}	π_{+2}	$\pi_{++}=1$

Modèle d'indépendance

- Si `Sexe` et `Chômage` ne sont pas liés (*indépendance*), alors

$$\pi_{ij} = \pi_{i+} \times \pi_{+j} \Rightarrow \log \mu_{ij} = \log(n_{++}\pi_{ij}) = \log n_{++} + \log \pi_{i+} + \log \pi_{+j}$$

- Sous l'hypothèse d'indépendance, nous pouvons donc écrire

$$\log \mu_{ij} = \gamma + \lambda_i^X + \lambda_j^Y \quad \text{avec} \quad \lambda_i^X = \log \pi_{i+} - \log \pi_{1+} \quad ; \quad \lambda_j^Y = \log \pi_{+j} - \log \pi_{+1}$$

Cela implique notamment que $\lambda_1^X = \lambda_1^Y = 0$ et $\log \mu_{11} = \gamma$.

Interprétation des paramètres. . . si indépendance *

- $\gamma = \log \mu_{11}$: c'est le log. du nombre attendu de personnes dans la cellule (1,1).
- $\lambda_i^X = \log\left(\frac{\pi_{i+}}{\pi_{1+}}\right) = \log\left(\frac{\mathrm{P}(X=i)}{\mathrm{P}(X=1)}\right) \quad ; \quad \lambda_j^Y = \log\left(\frac{\pi_{+j}}{\pi_{+1}}\right) = \log\left(\frac{\mathrm{P}(Y=j)}{\mathrm{P}(Y=1)}\right)$

On remarque que les catégories $X = 1$ et $Y = 1$ servent de référence.

Modèle d'indépendance: inférence

? Si le modèle d'indépendance est adéquat, quelles sont les valeurs plausibles pour les paramètres γ, λ_i^X et λ_j^Y?

• Pour répondre, nous devons spécifier les distributions a priori pour ces paramètres et être capable de calculer la probabilité d'observer la table de contingence étudiée pour des valeurs données des paramètres (càd la vraisemblance):

A priori: en l'absence d'information pertinente, un a priori avec grande variance peut être choisi.

Vraisemblance: nous pouvons supposer (voir Annexes A1-A3) que

$$N_{ij} \sim \text{Pois}(\mu_{ij} = n_{++}\pi_{ij}).$$

La vraisemblance est donc une fonction connue des paramètres d'intérêt.

• Le théorème de Bayes permet de déduire la distribution a posteriori des paramètres impliqués à partir de l'a priori et de la vraisemblance.

Nous pouvons en dériver un ensemble de valeurs plausibles pour les paramètres d'intérêt.

Modèle d'indépendance: sortie logicielle typique

• La plupart des logiciels statistiques permettent de faire de la *régression log-linéaire*, càd d'estimer à partir de données de comptage (tels qu'apparaissant dans une table de contingence) les paramètres intervenant de manière additive dans la description du **logarithme** de la moyenne d'une distribution de Poisson:

$$\log \mu_{ij} = \gamma + \lambda_i^X + \lambda_j^Y$$

• Les résultats proposés permettent de construire une **approximation** normale pour la distribution a posteriori de chacun des paramètres:

freq - Paramètres estimés (chomage2.sta)
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	6,164992	0,042593	20950,60	0,000000
Sexe	Homme	2	-0,046448	0,055656	0,70	0,403971
Sexe	Femme	3	0,000000			
Chomage	Oui	4	-0,943540	0,061949	231,98	0,000000
Chomage	Non	5	0,000000			
Echelle			1,000000	0,000000		

$$(\lambda_2^X|\text{donnees}) \sim \mathcal{N}(-0.046, 0.055^2) \quad ; \quad (\lambda_2^Y|\text{donnees}) \sim \mathcal{N}(-0.94, 0.062^2)$$

• Avant d'éventuellement exploiter ces résultats, il faut déterminer si l'hypothèse d'indépendance entre `Sexe` et `Chomage` est plausible.

Plausibilité de l'hypothèse d'indépendance

• La plausibilité de cette hypothèse peut être appréhendée en comparant chacune des fréquences observées n_{ij} avec la moyenne ajustée (*fréquence* attendue).

• Cette dernière, si aucune information a priori pertinente n'est utilisée, est égale à $n_{i+} \times n_{+j}/n = \exp(\hat{\gamma} + \hat{\lambda}_i^X + \hat{\lambda}_j^Y)$.

	Chômage		
Sexe	Non	Oui	Total
Femme	455 (475.8)	206 (185.2)	661
Homme	475 (454.2)	156 (176.8)	631
	930	362	1292

• Une "fréquence" attendue est la moyenne de la fréquence correspondante dans les tables de contingence obtenues si

$\longrightarrow$ on répète le sondage un grand nombre de fois,

$\longrightarrow$ l'hypothèse d'indépendance est exactement vérifiée au niveau de la population: ce n'est pratiquement jamais le cas.

• A 1ère vue, le nombre de chômeurs est surestimé (sous-estimé) chez les hommes (femmes) sous l'hypothèse d'indépendance.

• Reste à déterminer si les différences observées sont suffisamment grandes pour déclarer l'hypothèse d'indépendance trop réductrice...

Modèle saturé: définition

• Pour répondre à cette dernière question, il est utile de comparer les qualités prédictives de ce modèle avec celui n'imposant pas l'hypothèse d'indépendance.

• Si l'indépendance n'est plus postulée,

$$\pi_{ij} \neq \pi_{i+} \times \pi_{+j} \quad \text{et} \quad \log \mu_{ij} = \log(n_{++}\pi_{ij}) \neq \gamma + \lambda_i^X + \lambda_j^Y$$

• Cependant, nous pouvons écrire

$$\log \mu_{ij} = \gamma + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

avec les contraintes $\lambda_1^X = \lambda_1^Y = \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$

• Ce modèle fait apparaître autant de paramètres que de fréquences dans la table de contingence: il est *saturé*.

Cela se manifeste notamment par des fréquences attendues exactement égales aux fréquences observées:

$$\hat{\mu}_{ij} = \exp(\hat{\gamma} + \hat{\lambda}_i^X + \hat{\lambda}_j^Y + \hat{\lambda}_{ij}^{XY}) = n_{ij}$$

Modèle saturé: interprétation des paramètres

- La cote du chômage chez les femmes est, par définition,

$$\text{Cote}(\text{chomage}|\text{femme}) = \frac{\pi_{2|i=1}}{\pi_{1|i=1}}$$

Or, $\pi_{j|i} = \dfrac{\Pr(X = i \ \& \ Y = j)}{\Pr(X = i)} = \dfrac{\pi_{ij}}{\pi_{i+}}$. Donc

$$\log \text{Cote}(\text{chomage}|\text{femme}) = \log \frac{\pi_{12}}{\pi_{11}} = \log \frac{n_{++}\pi_{12}}{n_{++}\pi_{11}} = \log \frac{\mu_{12}}{\mu_{11}} = \lambda_2^Y .$$

- De même, le logarithme de la cote du chômage chez les hommes vaut

$$\log \frac{\pi_{22}}{\pi_{21}} = \log \frac{n_{++}\pi_{22}}{n_{++}\pi_{21}} = \log \frac{\mu_{22}}{\mu_{21}} = \lambda_2^Y + \lambda_{22}^{XY} .$$

- Par conséquent, le logarithme du rapport des cotes du chômage des hommes versus les femmes vaut

$$\log \text{RC} = \log \frac{\text{Cote}(\text{chomage}|\text{homme})}{\text{Cote}(\text{chomage}|\text{femme})} = \lambda_{22}^{XY}$$

Le paramètre λ_{22}^{XY} quantifie le lien entre les 2 variables binaires.

Modèle saturé: sortie logicielle typique

• Les résultats proposés permettent de construire une **approximation** normale pour la distribution a posteriori de chacun des paramètres du modèle saturé

$$\log \mu_{ij} = \gamma + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

freq - Paramètres estimés (chomage2.sta)
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	6,120297	0,046881	17043,41	0,000000
Sexe	Homme	2	0,043017	0,065598	0,43	0,511969
Sexe	Femme	3	0,000000			
Chomage	Oui	4	-0,792421	0,083977	89,04	0,000000
Chomage	Non	5	0,000000			
Sexe*Chomage	1	6	-0,321038	0,124771	6,62	0,010081
Sexe*Chomage	2	7	0,000000			
Sexe*Chomage	3	8	0,000000			
Sexe*Chomage	4	9	0,000000			
Echelle			1,000000	0,000000		

$$(\lambda_2^X|\text{donnees}) \sim \mathcal{N}(0.043, 0.066^2) \quad ; \quad (\lambda_2^Y|\text{donnees}) \sim \mathcal{N}(-0.79, 0.084^2)$$
$$(\lambda_{22}^{XY}|\text{donnees}) \sim \mathcal{N}(-0.32, 0.125^2)$$

• On peut vérifier que les fréquences attendues pour ce modèle saturé sont bien égales aux observées: par exemple,

$$\hat{\mu}_{12} = \exp(\hat{\gamma} + \hat{\lambda}_1^X + \hat{\lambda}_2^Y + \hat{\lambda}_{12}^{XY}) = \exp(6.120297 - 0.792421) = 206 = n_{12}$$

Modèle saturé: utilisation des résultats

• Il est maintenant possible d'évaluer si un lien entre `Sexe` et `Chomage` est détectable.

• Cela peut se faire via un intervalle de crédibilité 95% approximatif pour le log(RC) (déduit de sa distribution a posteriori), λ_{22}^{XY}:

$$-0.32 \pm 2 \times 0.125 = (-0.57, -0.07)$$

• Zéro ne faisant pas partie des valeurs retenues, cela suggère que l'indépendance entre `Sexe` et `Chomage` n'est pas une hypothèse plausible.

• La cote du chomage chez les hommes est estimée égale à $\mathrm{e}^{-0.32} = 73\% = 0.33/0.45$ de cette même cote chez les femmes, indiquant que le risque d'être chômeur est significativement plus faible pour les hommes.

	Chômage			
Sexe	Non	Oui	Cote	Total
Femme	455 (69%)	206 (31%)	0.45	661 (100%)
Homme	475 (75%)	156 (25%)	0.33	631 (100%)

• Notons que les valeurs comprises dans l'intervalle $(\mathrm{e}^{-0.57} = 57\%, \mathrm{e}^{-0.07} = 93\%)$ sont plausibles pour ce pourcentage. Un échantillon (ici, $n = 1292$) de plus grande taille permettrait d'affiner cette estimation.

Modèle saturé et test de rapport de vraisemblance

• Nous avons vu comment une approximation normale de l'a posteriori du log(RC) peut être utilisée pour déterminer si l'hypothèse d'indépendance n'est pas trop réductrice.

• Comparer les fréquences attendues sous le modèle d'indépendance avec celles du modèle relâchant cette hypothèse (ici le modèle saturé) est une façon moins approximative de **déterminer si 0 est une valeur plausible pour log(RC)**.

• Les fréquences attendues sous le modèle saturé étant exactement les fréquences observées, cela revient à comparer les fréquences attendues sous indépendance avec celles de la table de contingence de départ.

• Le *test de rapport de vraisemblance (ou chi-carré)* suggère de comparer ces valeurs:

$$G^2 = -2 \sum_i \sum_j n_{ij} \log \frac{A_{ij}}{n_{ij}}$$

où A_{ij} est la 'fréquence' attendue dans la cellule (i, j).

• Si $G^2 > \chi^2_1(0.95) = 3.84$, alors 0 ne fait pas partie des valeurs plausibles pour λ^{XY}_{22}, ($\chi^2_1(0.95)$ étant le quantile 95% de la distribution chi-carré avec un degré de liberté). Comme $G^2_{\text{obs}} = 6.66$, on conclut qu'il existe un lien significatif entre `Sexe` et `Chomage`.

Lien entre modèle log-linéaire et régression logistique

• Lorsqu'une des 2 variables intervenant dans la table de contingence est binaire et considérée comme la variable réponse (*dépendante*), il est possible de faire l'analyse par régression logistique:

$$\log \frac{\pi_i}{1-\pi_i} = \theta_i = \mu + \alpha_i \ ; \ (\alpha_1 = 0)$$

• On peut retrouver les résultats de la logistique via les estimations obtenues par le modèle log-linéaire.

Chomage - Paramètres estimés (chomage2.sta)
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	-0,792421	0,083977	89,04087	0,000000
Sexe	Homme	2	-0,321038	0,124771	6,62045	0,010081
Sexe	Femme	3	0,000000			
Echelle			1,000000	0,000000		

freq - Paramètres estimés (chomage2.sta)
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	6,120297	0,046881	17043,41	0,000000
Sexe	Homme	2	0,043017	0,065598	0,43	0,511969
Sexe	Femme	3	0,000000			
Chomage	Oui	4	-0,792421	0,083977	89,04	0,000000
Chomage	Non	5	0,000000			
Sexe*Chomage	1	6	-0,321038	0,124771	6,62	0,010081
Sexe*Chomage	2	7	0,000000			
Sexe*Chomage	3	8	0,000000			
Sexe*Chomage	4	9	0,000000			
Echelle			1,000000	0,000000		

• On retrouve, sans surprise le logarithme du rapport des cotes du chômage des hommes versus les femmes: c'est α_2 en régression logistique et λ_{22}^{XY} en régression log-linéaire.

Lien entre modèle log-linéaire et régression logistique (suite)

• L'intercept, μ, de la régression logistique se manifeste également via λ_2^Y.

Remarque

• La régression log-linéaire ne postule aucune variable réponse dans son analyse du lien entre `Sexe` et `Chomage`: les variables sont traitées symétriquement.

• Il est, en effet, tout aussi légitime d'étudier comment la valeur relative des proportions d'hommes et de femmes change d'un statut de non-chômeur à celui d'un chômeur.

• Les paramètres de la régression logistique correspondante sont alors λ_2^X et, comme précédemment, λ_{22}^{XY}.

Table $2 \times J$: modèle log-linéaire

• Ignorons pour l'instant l'effet `Sexe` et concentrons-nous sur le lien possible entre `Age.cat` et `Chomage`.

• La réponse est binaire et aboutit à une table de contingence $2 \times J$ [`chomage3.sta`].

	Age			
Chômage	30-39	40-49	50-65	65+
Non	184 (58%)	197 (61%)	304 (80%)	245 (91%)
Oui	131 (42%)	128 (39%)	78 (20%)	25 (9%)
Total	315 (100%)	325 (100%)	382 (100%)	270 (100%)

• Un modèle log-linéaire peut être considéré:

$$\log \mu_{ij} = \gamma + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

avec les contraintes $\lambda_1^X = \lambda_1^Y = \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$.

• Le paramètre λ_{2j}^{XY} a pour interprétation

$$\log \frac{\text{Cote}(\text{chomage}|\text{Age.cat } j)}{\text{Cote}(\text{chomage}|\text{Age.cat } 1)}$$

Table $2 \times J$: hypothèse d'indépendance

• Avant d'examiner les estimations obtenues pour les λ_{2j}^{XY}, il est recommandé d'évaluer globablement s'il est possible de mettre en évidence un lien entre `Chomage` et `Age.cat`, ce qui reviendrait à dire que la configuration $\lambda_{22}^{XY} = \ldots = \lambda_{24}^{XY} = 0$ n'est pas plausible.

• Comme dans l'exemple précédent, cela peut se faire par une comparaison des 'fréquences' attendues du modèle non contraignant avec celles du modèle d'indépendance via le test de rapport de vraisemblance:

$$G^2 = -2 \sum_i \sum_j n_{ij} \log \frac{A_{ij}}{n_{ij}}$$

où A_{ij} est la 'fréquence' attendue dans la cellule (i, j).

• Si $G^2 > \chi^2_{J-1=3}(0.95) = 7.81$, alors $\lambda_{22}^{XY} = \ldots = \lambda_{24}^{XY} = 0$ est une configuration des 4 logarithmes de rapport de cotes n'appartenant pas à la région de crédibilité 95% pour $(\lambda_{22}^{XY}, \ldots, \lambda_{24}^{XY})$.

L'hypothèse d'indépendance sera alors déclarée *non plausible* et l'interaction entre X et Y *significative*.

Table $2 \times J$: hypothèse d'indépendance (suite)

• **Note**: le nombre de degrés de liberté $J - 1$ de la chi-carré correspond au nombre de paramètres contraints dans le modèle d'intérêt.

• Comme $G^2_{\text{obs}} = 115.85 > 7.81$, on conclut qu'il existe un lien significatif entre `Chomage` et `Age`.

freq - Test Vraisemblance Type 3 (chomage3.sta)
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Age.cat	3	-53,284	52,8305	0,000000
Chomage	1	-172,261	290,7859	0,000000
Age.cat*Chomage	3	-84,792	115,8480	0,000000

• On aboutit à la même conclusion en constatant que la P-valeur associée (0.00) est inférieure à 5%.

• L'examen des estimations pour les λ^{XY}_{2j} va nous permettre de comprendre la nature du lien entre `Chomage` et `Age.cat`.

Table $2 \times J$: sortie logicielle typique

- Voici les estimations pour les paramètres de régression logistique et log-linéaire:

Chomage - Paramètres estimés (chomage3.sta)
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	-2,28238	0,209955	118,1750	0,000000
Age.cat	30-39	2	1,94264	0,239059	66,0352	0,000000
Age.cat	40-49	3	1,85121	0,238683	60,1544	0,000000
Age.cat	50-64	4	0,92206	0,245338	14,1251	0,000171
Age.cat	65 et +	5	0,00000			
Echelle			1,00000	0,000000		

freq - Paramètres estimés (chomage3.sta)
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	5,50126	0,063888	7414,641	0,000000
Age.cat	30-39	2	-0,28632	0,097552	8,615	0,003335
Age.cat	40-49	3	-0,21805	0,095696	5,192	0,022690
Age.cat	50-64	4	0,21577	0,085855	6,316	0,011965
Age.cat	65 et +	5	0,00000			
Chomage	Oui	6	-2,28238	0,209956	118,173	0,000000
Chomage	Non	7	0,00000			
Age.cat*Chomage	1	8	1,94264	0,239061	66,034	0,000000
Age.cat*Chomage	2	9	0,00000			
Age.cat*Chomage	3	10	1,85121	0,238684	60,154	0,000000
Age.cat*Chomage	4	11	0,00000			
Age.cat*Chomage	5	12	0,92206	0,245340	14,125	0,000171
Age.cat*Chomage	6	13	0,00000			
Age.cat*Chomage	7	14	0,00000			
Age.cat*Chomage	8	15	0,00000			
Echelle			1,00000	0,000000		

Les catégories de référence choisies par le logiciel sont 'Non' pour `Chomage` et '65+' pour `Age.cat`.

- Les λ_{2j}^{XY} (et leur erreurs standards) sont estimés par:
  -30-39 vs. 65+: 1.94 (0.239) ; 40-49 vs. 65+: 1.85 (0.239)
  -50-64 vs. 65+: 0.92 (0.245).

Table $2 \times J$: sortie logicielle typique (suite)

Etiquettes des Colonnes (chomage3.sta)
Etiquettes des colonnes de la matrice X du modèle

Etiq.	Colonne	Variable	Niveau Variable	Variable	Niveau Variable
Ord.Orig.	1				
Age.cat	2	Age.cat	30-39		
Age.cat	3	Age.cat	40-49		
Age.cat	4	Age.cat	50-64		
Age.cat	5	Age.cat	65 et +		
Chomage	6	Chomage	Oui		
Chomage	7	Chomage	Non		
Age.cat*Chomage	8	Age.cat	30-39	Chomage	Oui
Age.cat*Chomage	9	Age.cat	30-39	Chomage	Non
Age.cat*Chomage	10	Age.cat	40-49	Chomage	Oui
Age.cat*Chomage	11	Age.cat	40-49	Chomage	Non
Age.cat*Chomage	12	Age.cat	50-64	Chomage	Oui
Age.cat*Chomage	13	Age.cat	50-64	Chomage	Non
Age.cat*Chomage	14	Age.cat	65 et +	Chomage	Oui
Age.cat*Chomage	15	Age.cat	65 et +	Chomage	Non

freq - Intervalles de Conf. des Estimations
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	LC Inf. 95, %	LC Sup. 95, %
Ord.Orig		1	5,37604	5,62648
Age.cat	30-39	2	-0,47752	-0,09512
Age.cat	40-49	3	-0,40562	-0,03049
Age.cat	50-64	4	0,04750	0,38404
Age.cat	65 et +	5		
Chomage	Oui	6	-2,69389	-1,87088
Chomage	Non	7		
Age.cat*Chomage	1	8	1,47409	2,41119
Age.cat*Chomage	2	9	-0,00000	0,00000
Age.cat*Chomage	3	10	1,38340	2,31902
Age.cat*Chomage	4	11	-0,00000	0,00000
Age.cat*Chomage	5	12	0,44121	1,40292
Age.cat*Chomage	6	13		
Age.cat*Chomage	7	14	-0,00000	0,00000
Age.cat*Chomage	8	15	-0,00000	0,00000

• Les intervalles de crédibilité 95% pour les λ_{2j}^{XY} correspondant au 3 premières catégories d'âge ne contiennent que des valeurs positives, ce qui suggère un risque d'avoir été 'chômeur' plus élevé que chez les 65 ans et plus.

• Les catégories '30-39' et '40-49' sont les plus exposées, la catégorie '50-64' connaissant un risque intermédiaire.

Table $I \times J \times K$: point de départ

• Dans notre exemple, il serait intéressant d'intégrer les variables `Chomage` (X), `Sexe` (Y) et `Age.cat` (Z) dans un même modèle pour, notamment, évaluer les liens conjoints des 2 dernières variables avec la 1ère.

• Le point de départ est la table de contigence suivante:

		Age			
Sexe	Chômage	30-39	40-49	50-65	65+
Homme	Oui	59	52	31	14
	Non	113	106	146	110
Femme	Oui	72	76	47	11
	Non	71	91	158	135

• On peut en déduire (les estimations de) la cote du chômage pour les $4 \times 2 = 8$ combinaisons de `Age` et `Sexe`:

	Age			
Sexe	30-39	40-49	50-65	65+
Homme	0.52	0.49	0.21	0.13
Femme	1.01	0.84	0.30	0.08

• La cote du chômage semble diminuer (de la même manière?) avec l'âge chez les hommes et les femmes. Dans une catégorie d'âge donnée, la cote du chômage tend à être plus élevée chez les femmes que chez les hommes.

Table $I \times J \times K$: modèle d'indépendance

- Notations:
  -$\pi_{ijk} = \Pr(X = i, Y = j, Z = k)$.
  -n_{ijk}: fréquence associée à la cellule (i, j, k).
  -$\mu_{ijk} = n_{+++}\pi_{ijk}$.

- Sous l'hypothèse d'indépendance entre X et les 2 autres variables,

$$\pi_{ijk} = \pi_{i+} \times \pi_{+jk} \Rightarrow \log \mu_{ijk} = \log(n_{+++}\pi_{ijk}) = \log n_{+++} + \log \pi_{i+} + \log \pi_{+jk}$$

- Sous cette hypothèse, nous pouvons donc écrire

$$\log \mu_{ijk} = \gamma + \lambda_i^X + (\lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ})$$

avec les contraintes d'identification habituelles (0 pour un coefficient dès que 1 apparaît dans les indices).

- Ecriture synthétique pour ce modèle: `X+(Y+Z+YZ)`.

Table $I \times J \times K$: modèle d'indépendance (2)

Remarques

• Ces contraintes impliquent que la 1ère catégorie est désignée comme référence pour chacune des variables.

• Les termes $(\lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ})$ indiquent qu'aucune contrainte n'est imposée pour décrire π_{+jk}.

• Comme prédédemment, la pertinence de l'hypothèse simplificatrice d'indépendance peut être évaluée en comparant ses 'fréquences' attendues avec celles sous un modèle moins contraignant.

Table $I \times J \times K$: modèle saturé

• Le modèle *saturé* n'impose aucune structure de dépendance particulière pour (X, Y, Z):

$$\log \mu_{ijk} = \gamma + \lambda_i^X + (\lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}) + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$$

avec les contraintes d'identification habituelles.

• Ecriture synthétique pour ce modèle: `X+(Y+Z+YZ)+XY+XZ+XYZ`.

• Ce modèle contient autant de paramètres que de fréquences observées: il ne résume donc pas l'information disponible.

• Les 'fréquences' attendues pour ce modèle sont exactement également aux fréquences observées.

Table $I \times J \times K$: coefficients du modèle saturé

• Focalisons-nous sur l'interpréation des coefficients quantifiant les liens entre X et les variables Y et Z.

• Définition: $\text{Cote}(X = i|Y = j, Z = k) = \dfrac{\text{P}(X = i|Y = j, Z = k)}{\text{P}(X = 1|Y = j, Z = k)}$.

• On peut montrer que

$$\lambda_{ij}^{XY} = \log \frac{\text{Cote}(X = i|Y = j, Z = 1)}{\text{Cote}(X = i|Y = 1, Z = 1)}$$

$$\lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ} = \log \frac{\text{Cote}(X = i|Y = j, Z = k)}{\text{Cote}(X = i|Y = 1, Z = k)}$$

$$\lambda_{ik}^{XZ} = \log \frac{\text{Cote}(X = i|Y = 1, Z = k)}{\text{Cote}(X = i|Y = 1, Z = 1)}$$

$$\lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ} = \log \frac{\text{Cote}(X = i|Y = j, Z = k)}{\text{Cote}(X = i|Y = j, Z = 1)}$$

Table $I \times J \times K$: hypothèses d'intérêt

• Plusieurs hypothèses simplificatrices peuvent être émises à partir du modèle saturé: elles correspondent à la nullité de certains paramètres.

• La plausibilité de ces hypothèses peut être évaluée par un test de rapport de vraisemblance comparant les 'fréquences' attendues sous le modèle correspondant avec les fréquences observées (qui sont les fréquences attendues du modèle saturé) ou les 'fréquences' attendues du dernier modèle retenu.

Pas d'interaction du 3ème ordre: `X+(Y+Z+YZ)+XY+XZ`

• Hypothèse: $\lambda_{ijk}^{XYZ} = 0$, càd $(I-1)(J-1)(K-1)$ paramètres contraints.

• Cela implique que le lien entre X et Y (ou X et Z) n'est pas affecté par Z (Y).

• Le test de rapport de vraisemblance comparant ce modèle au saturé est

$$G^2 = -2 \sum_i \sum_j \sum_k n_{ijk} \log \frac{A_{ijk}}{n_{ijk}}$$

Le nombre de degrés de liberté associé est $(I-1)(J-1)(K-1)$.

Table $I \times J \times K$: hypothèses d'intérêt (2)

Indépendance conditionnelle de X & Z étant donné Y: $(X \perp Z|Y)$

- Notation: `X+(Y+Z+YZ)+XY`.
- Cette hypothèse signifie que

$$\mathrm{P}(X = i, Z = k|Y = j) = \mathrm{P}(X = i|Y = j)\mathrm{P}(Z = k|Y = j)$$

et implique que $\lambda_{ik}^{XZ} = 0$, soient $(I-1)(K-1)$ paramètres contraints.

- **Note**: il n'est pertinent d'évaluer cette hypothèse que si l'absence d'interaction du 3ème ordre est plausible.
- Le test de rapport de vraisemblance comparant les 'fréquences' attendues $\{A_{ijk}\}$ sous ce modèle à celles $\{A_{ijk}^{ref}\}$ du modèle sans interaction du 3ème ordre est

$$\Delta G^2 = G^2(A_{ijk}) - G^2(A_{ijk}^{ref}) = -2\sum_i \sum_j \sum_k n_{ijk} \log \frac{A_{ijk}}{A_{ijk}^{ref}}.$$

Le nombre de degrés de liberté associé est $(I-1)(K-1)$.

Table $I \times J \times K$: hypothèses d'intérêt (3)

Indépendance conditionnelle de X & Y étant donné Z: $(X \perp Y|Z)$

- Notation: `X+(Y+Z+YZ)+XZ`.
- Cette hypothèse signifie que

$$\mathrm{P}(X=i, Y=j|Z=k) = \mathrm{P}(X=i|Z=k)\mathrm{P}(Y=j|Z=k)$$

et implique que $\lambda_{ij}^{XY} = 0$, soient $(I-1)(J-1)$ paramètres contraints.

- **Note**: il n'est pertinent d'évaluer cette hypothèse que si l'absence d'interaction du 3ème ordre est plausible.
- Le test de rapport de vraisemblance comparant les 'fréquences' attendues $\{A_{ijk}\}$ sous ce modèle à celles $\{A_{ijk}^{ref}\}$ du modèle sans interaction du 3ème ordre est

$$\Delta G^2 = G^2(A_{ijk}) - G^2(A_{ijk}^{ref}) = -2\sum_i \sum_j \sum_k n_{ijk} \log \frac{A_{ijk}}{A_{ijk}^{ref}}.$$

Le nombre de degrés de liberté associé est $(I-1)(J-1)$.

Table $2 \times J \times K$: exemple

• Considérons les modèles précédents dans notre exemple pour étudier les liens entre `Chomage` (X) et les variables `Sexe` (Y), `Age.cat` (Z).

• La stratégie de modélisation proposée consiste à démarrer avec le modèle saturé `X+(Y+Z+YZ)+XY+XZ+XYZ` et à progressivement l'amputer des termes d'interaction impliquant X.

• Les 2 tables suivantes vont nous être utiles:

Freq - Test Vraisemblance Type 1 (chomag
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-256,733		
Sexe	1	-256,385	0,6967	0,403909
Age.cat	3	-246,584	19,6027	0,000205
Chomage	1	-117,360	258,4478	0,000000
Sexe*Age.cat	3	-114,322	6,0750	0,108017
Sexe*Chomage	1	-110,991	6,6617	0,009851
Age.cat*Chomage	3	-50,950	120,0822	0,000000
Sexe*Age.cat*Chomage	3	-48,119	5,6622	0,129253

Freq - Test Vraisemblance Type 1 (chomag
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-256,733		
Chomage	1	-127,509	258,4478	0,000000
Sexe	1	-127,161	0,6967	0,403909
Age.cat	3	-117,360	19,6027	0,000205
Sexe*Age.cat	3	-114,322	6,0750	0,108017
Age.cat*Chomage	3	-56,398	115,8480	0,000000
Sexe*Chomage	1	-50,950	10,8959	0,000964
Sexe*Age.cat*Chomage	3	-48,119	5,6622	0,129253

Table $2 \times J \times K$: simplication du modèle

Retrait de XYZ: pas d'interaction du 3ème ordre?

• Le terme d'ordre 3, XYZ, est enlevé le premier. Ce retrait est acceptable pour autant que la configuration $\lambda_{ijk}^{XYZ} = 0 \; \forall i, j, k$ soit plausible.

• La table des tests de vraisemblance de type I (correspondant à un retrait séquentiel des termes indiqués du modèle saturé) nous indique que $G^2 = 5.66$ pour les 3 paramètres ainsi contraints. La P-valeur associée, $0.129 > 0.05$, nous apprend que la simplification envisagée est plausible.

• Conclusion: l'interaction entre Chomage et Sexe (Age.cat), si elle existe, ne dépend pas de Age.cat (Sexe).

	Age			
Sexe	30-39	40-49	50-65	65+
Homme	0.52	0.49	0.21	0.13
Femme	1.01	0.84	0.30	0.08
RC(H vs F)	0.53	0.58	0.70	1.63

Autrement dit, il n'y a pas d'indication que les estimations dans le tableau (reprenant les cotes du chômage) soient celles de rapports de cotes différents. On peut donc mesurer ce lien éventuel par un seul nombre commun aux 4 catégories d'âge.

Table $2 \times J \times K$: simplication du modèle (2)

Retrait de XY: a-t-on $(X \perp Y|Z)$*?*

• Peut-on enlever le terme XY du modèle déjà amputé de l'interaction du 3ème ordre? Ce retrait est acceptable pour autant que la configuration $\lambda_{ij}^{XY} = 0 \; \forall i, j$ soit plausible.

Cela reviendrait à dire que 1 est une valeur plausible pour le RC (que l'on peut désormais supposer) commun à chaque catégorie d'âge du tableau précédent.

• La table de droite (voir p. 30) nous indique que $\Delta G^2 = 10.90$ pour le seul paramètre ainsi contraint. La P-valeur associée, $0.00096 < 0.05$, nous apprend que la simplification envisagée n'est pas plausible.

• Conclusion: même après avoir tenu compte de l'effet Age (cf. terme XZ), il existe un lien significatif entre Chomage (X) et Sexe (Y).

• Le RC commun n'a donc pas 1 comme valeur plausible.

• La nature du lien sera examinée plus bas.

Table $2 \times J \times K$: simplication du modèle (3)

Retrait de XZ: a-t-on $(X \perp Z|Y)$?

• Peut-on enlever le terme XZ du modèle déjà amputé de l'interaction du 3ème ordre? Ce retrait est acceptable pour autant que la configuration $\lambda_{ik}^{XZ} = 0 \ \forall(i,k)$ soit plausible.

• La table de gauche (voir p. 30) nous indique que $\Delta G^2 = 120.08$ pour les 3 paramètres ainsi contraints. La P-valeur associée, $0.000 < 0.05$, nous apprend que la simplification envisagée n'est pas plausible.

• Conclusion: même après avoir tenu compte d'un effet Sexe (cf. terme XY) éventuel, il existe un lien significatif entre Chomage (X) et Age.cat (Z). La nature de ce lien sera examinée plus bas.

Table $2 \times J \times K$: estimations dans le modèle final

• Le modèle retenu est donc `X+(Y+Z+YZ)+XY+XZ`. Il indique que le risque d'être chômeur dépend du sexe et de l'âge, l'effet 'sexe' ('âge') ne dépendant pas de l'âge (du sexe) des personnes.

• L'examen des coefficients relatifs à `XY` et `XZ` dans ce modèle va nous permettre de comprendre la nature de ces liens.

Etiquettes des Colonnes (chomage.sta)
Etiquettes des colonnes de la matrice X du modèle

Etiq.	Colonne	Variable	Niveau Variable	Variable	Niveau Variable
Age.cat*Chomage	18	Age.cat	30-39	Chomage	OUI
Age.cat*Chomage	19	Age.cat	30-39	Chomage	NON
Age.cat*Chomage	20	Age.cat	40-49	Chomage	OUI
Age.cat*Chomage	21	Age.cat	40-49	Chomage	NON
Age.cat*Chomage	22	Age.cat	49-64	Chomage	OUI
Age.cat*Chomage	23	Age.cat	49-64	Chomage	NON
Age.cat*Chomage	24	Age.cat	65+	Chomage	OUI
Age.cat*Chomage	25	Age.cat	65+	Chomage	NON
Sexe*Chomage	26	Sexe	HOMME	Chomage	OUI
Sexe*Chomage	27	Sexe	HOMME	Chomage	NON
Sexe*Chomage	28	Sexe	FEMME	Chomage	OUI
Sexe*Chomage	29	Sexe	FEMME	Chomage	NON

Freq - Paramètres estimés (chomage.sta)
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Age.cat*Chomage	1	18	1,99489	0,240574	68,761	0,000000
Age.cat*Chomage	2	19	0,00000			
Age.cat*Chomage	3	20	1,87641	0,239610	61,326	0,000000
Age.cat*Chomage	4	21	0,00000			
Age.cat*Chomage	5	22	0,92876	0,245935	14,262	0,000159
Age.cat*Chomage	6	23	0,00000			
Age.cat*Chomage	7	24	0,00000			
Age.cat*Chomage	8	25	0,00000			
Sexe*Chomage	1	26	-0,43102	0,131293	10,777	0,001028
Sexe*Chomage	2	27	0,00000			
Sexe*Chomage	3	28	0,00000			
Sexe*Chomage	4	29	0,00000			
Echelle			1,00000	0,000000		

Table $2 \times J \times K$: estimations dans le modèle final (2)

Interaction entre Chomage (X) et Sexe (Y) à un Age (Z) donné

- Les catégories de référence sont 'Non' pour Chomage et 'Femme' pour Sexe.
- $\lambda_{22}^{XY} \overset{\text{approx.}}{\sim} \mathcal{N}(-0.43, 0.131^2)$.
- Nous savons que

$$\lambda_{ij}^{XY} = \log \frac{\text{Cote}(X = i|Y = j, Z = 1)}{\text{Cote}(X = i|Y = 1, Z = 1)}$$
$$\lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ} = \log \frac{\text{Cote}(X = i|Y = j, Z = k)}{\text{Cote}(X = i|Y = 1, Z = k)}$$

- Comme les termes d'interaction du 3ème ordre sont supposés nuls, on peut conclure que, quelle que soit la catégorie d'âge (cf. Z), la cote du chômage ($X = 2$) chez les hommes ($Y = 2$) est estimée à $\text{e}^{-0.43} = 65\%$ de cette même cote chez les femmes ($Y = 1$).
- Un intervalle de crédibilité 95% pour ce pourcentage est donné par $\text{e}^{-0.43 \pm 2 \times 0.131} = (50\%, 85\%)$.

Table $2 \times J \times K$: estimations dans le modèle final (3)

• Ce lien entre `Chomage` et `Sexe` (pour une catégorie d'âge donnée) peut être mis en évidence à partir des estimations des rapports de cotes du chômage des hommes vs les femmes:

	Age			
Sexe	30-39	40-49	50-65	65+
Homme	0.52	0.49	0.21	0.13
Femme	1.01	0.84	0.30	0.08
RC(H vs F)	0.53	0.58	0.70	1.63

• Rappel: l'analyse qui a précédé nous indique que ces 4 rapports de cotes ne sont pas significativement différents (cfr. pas d'interaction du 3ème ordre significative).

Il est donc légitime d'estimer l'effet Sexe à un âge donné par une seule quantité $\exp(\hat{\lambda}_{22}^{XY}) = \exp(-0.43) = 0.65$. C'est de toute évidence un compromis entre les 4 rapports de cotes du tableau.

• Nous savons également que 1 n'est pas une valeur plausible pour ce rapport de cotes: la cote du chômage est donc significativement inférieure chez les hommes d'une catégorie d'âge donnée.

Table $2 \times J \times K$: estimations dans le modèle final (4)

Interaction entre `Chomage` (X) et `Age.cat` (Z) pour un `Sexe` (Y) donné

• Les catégories de référence sont 'Non' pour `Chomage` et '65+' pour `Age.cat`. Numérotons les catégories d'âge par $k = 1$: 65+ ; $k = 2$: 50-64 ; $k = 3$: 40-49 ; $k = 4$: 30-39.

• $\lambda_{22}^{XZ} \overset{\text{approx.}}{\sim} \mathcal{N}(0.93, 0.246^2)$; $\lambda_{23}^{XZ} \overset{\text{approx.}}{\sim} \mathcal{N}(1.88, 0.240^2)$; $\lambda_{24}^{XZ} \overset{\text{approx.}}{\sim} \mathcal{N}(1.99, 0.241^2)$.

• Nous savons que

$$\lambda_{ik}^{XZ} = \log \frac{\text{Cote}(X = i | Y = 1, Z = k)}{\text{Cote}(X = i | Y = 1, Z = 1)}$$

$$\lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ} = \log \frac{\text{Cote}(X = i | Y = j, Z = k)}{\text{Cote}(X = i | Y = j, Z = 1)}$$

• Comme les termes d'interaction du 3ème ordre sont supposés nuls, on peut conclure que, quel que soit le sexe (cf. Y), la cote du chômage ($X = 2$) chez les 50-64 ans (resp. 40-49, 30-39) ($Y = 2$) est estimée $\mathrm{e}^{0.93} = 2.53$ (resp. 6.55, 7.31) plus grande que cette même cote chez les 65 ans et plus ($Z = 1$).

Table $2 \times J \times K$: estimations dans le modèle final (5)

• L'intervalle de crédibilité 95% pour ce facteur de proportionalité chez les 50-64 (resp. 40-49, 30-39) ans est donné par $\mathrm{e}^{0.93 \pm 2 \times 0.246} = (1.55, 4.15)$ (resp. $(4.06, 10.6)$, $(4.52, 11.8)$).

• On peut estimer les cotes du chômage (les rapports de cotes du chômage d'une catégorie d'âge donnée vs les 65+ pour sexe donné) directement à partir des données:

	Age			
Sexe	30-39	40-49	50-65	65+
Homme	0.52 (4.00)	0.49 (3.77)	0.21 (1.62)	0.13 (1.00)
Femme	1.01 (12.6)	0.84 (10.5)	0.30 (3.75)	0.08 (1.00)

• L'analyse qui a précédé nous indique qu'il n'est pas nécessaire d'estimer séparément chez les hommes et les femmes les rapports de cotes contrastant une catégorie d'âge donnée avec les 65+.

• Les estimations fournies par le modèle sont rappelées ci-contre:

Age	30-39	40-49	50-65	65+
RC	7.31	6.55	2.53	1.00

C'est de tout évidence un compromis entre les RC pour les hommes et les femmes.

Table $2 \times J \times K$: estimations dans le modèle final (6)

• Conclusion: le risque d'être 'chômeur' s'est significativement accru au cours du temps avec un accroissement déjà perceptible chez les 50-64 ans et des risques de magnitudes semblables chez les 30-39 et 40-49 ans.

Remarquons cependant que le temps passé sur le marché du travail étant plus faible chez les 30-39 que chez les 40-49 ans, il est vraisemblable que la cohorte des 30-39 ans soit finalement (à l'heure du bilan en fin de carrière) bien plus encore sujette à des expériences de chômage que la cohorte des 40-49 ans.

Table $2 \times J \times K$: régression logistique

• Lorsque la variable réponse est binaire, il est possible de procéder à la même analyse par régression logistique.

• Le modèle prend alors la forme

$$\log \frac{\pi_{jk}}{1-\pi_{jk}} = \theta_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}$$
$$(\alpha_1 = \beta_1 = \gamma_{1k} = \gamma_{j1} = 0)$$

• Si $j = 1$ est la catégorie 'Femme' pour `Sexe` et $k = 1$ la catégorie '65+' pour `Age.cat`:

$$\alpha_2 = \log \frac{\text{Cote(chomage|Homme et 65+)}}{\text{Cote(chomage|Femme et 65+)}}$$
$$\beta_k = \log \frac{\text{Cote(chomage|Femme et Age.cat k)}}{\text{Cote(chomage|Femme et 65+)}}$$
$$\alpha_2 + \gamma_{2k} = \log \frac{\text{Cote(chomage|Homme et Age.cat k)}}{\text{Cote(chomage|Femme et Age.cat k)}}$$
$$\beta_k + \gamma_{2k} = \log \frac{\text{Cote(chomage|Homme et Age.cat k)}}{\text{Cote(chomage|Homme et 65+)}}$$

Table $2 \times J \times K$: régression logistique (2)

• Nous pouvons donc faire les correspondances suivantes entre les paramètres de régression logistique et log-linéaire:

$$\alpha_j = \lambda_{2j}^{XY} \ ; \ \beta_k = \lambda_{2k}^{XZ} \ ; \ \gamma_{jk} = \lambda_{2jk}^{XYZ}$$

• Bien entendu, une même stratégie de modélisation conduit exactement au même modèle final et aux mêmes interprétations et estimations.

• Les estimations obtenues pour le modèle sans interaction peuvent être comparées et constatées identiques en logistique et en log-linéaire:

Chomage - Paramètres estimés (chomage.sta)
Distribution : BINOMIALE
Fonction de Liaison : LOGIT

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Ord.Orig		1	-2,10302	0,215987	94,80563	0,000000
Sexe	HOMME	2	-0,43102	0,131293	10,77730	0,001028
Sexe	FEMME	3	0,00000			
Age.cat	30-39	4	1,99489	0,240573	68,76121	0,000000
Age.cat	40-49	5	1,87641	0,239609	61,32663	0,000000
Age.cat	49-64	6	0,92876	0,245934	14,26180	0,000159
Age.cat	65+	7	0,00000			
Echelle			1,00000	0,000000		

Freq - Paramètres estimés (chomage.sta)
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Age.cat*Chomage	1	18	1,99489	0,240574	68,761	0,000000
Age.cat*Chomage	2	19	0,00000			
Age.cat*Chomage	3	20	1,87641	0,239610	61,326	0,000000
Age.cat*Chomage	4	21	0,00000			
Age.cat*Chomage	5	22	0,92876	0,245935	14,262	0,000159
Age.cat*Chomage	6	23	0,00000			
Age.cat*Chomage	7	24	0,00000			
Age.cat*Chomage	8	25	0,00000			
Sexe*Chomage	1	26	-0,43102	0,131293	10,777	0,001028
Sexe*Chomage	2	27	0,00000			
Sexe*Chomage	3	28	0,00000			
Sexe*Chomage	4	29	0,00000			
Echelle			1,00000	0,000000		

Réponse polytomiale

• Lorsque la variable réponse est polytomiale, il n'est plus possible de faire une régression logistique binaire.

Ex European Social Survey (année 2002 pour la Belgique) - Intéressons-nous à l'affirmation suivante: **Les homosexuels hommes et femmes devraient être libres de vivre leur vie comme ils le souhaitent.**.

Réponse possibles : 1: Tout à fait d'accord ; 2: Plutôt d'accord ; 3: Ni d'accord, ni en désaccord ; 4: Plutôt en désaccord ; 5: Tout à fait en désaccord.

• Comment la nature de la réaction (X) change-t-elle avec le sexe (Y) et la catégorie d'âge (Z) de la personne questionnée?

Réponse polytomiale (2)

• La table de contigence suivante peut être construite ($n = 1764$) [`homo.sta`]:

	Réaction, femmes					Réaction, hommes				
Age	1	2	3	4	5	1	2	3	4	5
< 30	129	85	11	7	5	109	70	34	18	10
$30 - 39$	70	55	7	6	4	79	59	20	6	8
$40 - 49$	82	59	9	12	8	66	53	18	13	8
$50 - 64$	94	72	18	7	12	66	76	19	8	8
> 65	47	45	20	15	17	38	45	15	8	14

• Si nous prenons la catégorie 1 comme référence pour `Réaction` ($X = 1$), la cote de la réponse $X = i$ pour une catégorie d'`Age` ($Y = j$) et de Sexe ($Z = k$) donnée est alors définie par

$$\text{Cote}(X = i|Y = j, Z = k) = \frac{\text{P}(X = i|Y = j, Z = k)}{\text{P}(X = 1|Y = j, Z = k)}$$

Réponse polytomiale (3)

• Ces cotes peuvent facilement être estimées à partir de la table de départ. **Ex** Cote de la réaction 2 ('Plutôt d'accord') chez les femmes de moins de 30 ans: $85/129 = 0.66$.

• On obtient ainsi les estimations suivantes pour les cotes:

	Réaction, femmes					Réaction, hommes				
Age	1	2	3	4	5	1	2	3	4	5
< 30	1.00	0.66	0.09	0.05	0.04	1.00	0.64	0.31	0.17	0.09
$30 - 39$	1.00	0.79	0.10	0.09	0.06	1.00	0.75	0.25	0.08	0.10
$40 - 49$	1.00	0.72	0.11	0.15	0.10	1.00	0.80	0.27	0.20	0.12
$50 - 64$	1.00	0.77	0.19	0.07	0.13	1.00	1.15	0.29	0.12	0.12
> 65	1.00	0.96	0.43	0.32	0.36	1.00	1.18	0.39	0.21	0.37

• Le modèle log-linéaire, tel que présenté précédemment, peut-être utilisé sans modification pour répondre à la question initiale.

• Le modèle retenu va permettre de déterminer ce qui structure (significativement) les cotes.

Réponse polytomiale: sélection du modèle

- Notation: X=Réaction ; Y=Age ; Z=Sexe.
- Le modèle d'indépendance est `X+(Y+Z+YZ)`.
- Le modèle de départ (ci-dessous) est le saturé `X+(Y+Z+YZ)+XY+XZ+XYZ`.

Freq - Test Vraisemblance Type 1 (homo.s
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-802,413		
Reaction	4	-209,806	1185,215	0,000000
Age	4	-173,448	72,715	0,000000
Sexe	1	-173,226	0,444	0,504976
Age*Sexe	4	-169,794	6,864	0,143256
Reaction*Age	16	-143,247	53,095	0,000007
Reaction*Sexe	4	-134,633	17,227	0,001746
Reaction*Age*Sexe	16	-123,922	21,422	0,162855

Freq - Test Vraisemblance Type 1 (homo.s
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Degré de Liberté	Log-Vraisbnc	Chi²	p
Ord.Orig.	1	-802,413		
Reaction	4	-209,806	1185,215	0,000000
Age	4	-173,448	72,715	0,000000
Sexe	1	-173,226	0,444	0,504976
Age*Sexe	4	-169,794	6,864	0,143256
Reaction*Sexe	4	-162,086	15,415	0,003913
Reaction*Age	16	-134,633	54,907	0,000004
Reaction*Age*Sexe	16	-123,922	21,422	0,162855

- L'interaction du 3ème ordre n'est pas significative.
- Il existe une interaction significative entre X et les 2 variables explicatives Y et Z.
- Le modèle retenu est donc `X+(Y+Z+YZ)+XY+XZ`.

Réponse polytomiale: interprétation

Freq - Paramètres estimés (homo.sta)
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Reaction*Sexe	1	24	-0,27756	0,221465	1,571	0,210106
Reaction*Sexe	2	25	0,00000			
Reaction*Sexe	3	26	-0,32749	0,214365	2,334	0,126585
Reaction*Sexe	4	27	0,00000			
Reaction*Sexe	5	28	-0,68699	0,174281	15,538	0,000081
Reaction*Sexe	6	29	0,00000			
Reaction*Sexe	7	30	-0,13883	0,108403	1,640	0,200313
Reaction*Sexe	8	31	0,00000			
Reaction*Sexe	9	32	0,00000			
Reaction*Sexe	10	33	0,00000			

Etiquettes des Colonnes (homo.sta)
Etiquettes des colonnes de la matrice X du modèle

Etiq.	Colonne	Variable	Niveau Variable	Variable	Niveau Variable
Reaction*Sexe	24	Reaction	5	Sexe	Femme
Reaction*Sexe	25	Reaction	5	Sexe	Homme
Reaction*Sexe	26	Reaction	4	Sexe	Femme
Reaction*Sexe	27	Reaction	4	Sexe	Homme
Reaction*Sexe	28	Reaction	3	Sexe	Femme
Reaction*Sexe	29	Reaction	3	Sexe	Homme
Reaction*Sexe	30	Reaction	2	Sexe	Femme
Reaction*Sexe	31	Reaction	2	Sexe	Homme
Reaction*Sexe	32	Reaction	1	Sexe	Femme
Reaction*Sexe	33	Reaction	1	Sexe	Homme

• La catégorie de référence pour `Réaction` est 1 ($X = 1$: 'Tout à fait d'accord') et celle pour `Sexe` est 'Homme' ($Z = 1$).

Note: nous avons forcé la catégorie 1 comme référence pour `Réaction` car c'est la plus populaire (cf. variance des estimateurs des logarithmes de rapport de cotes).

• Puisque $\lambda_{ijk}^{XYZ} = 0$, les estimations réfèrent à

$$\lambda_{ik}^{XZ} = \log \frac{\text{Cote}(X = i|Y = j, Z = k)}{\text{Cote}(X = i|Y = j, Z = 1)} \quad \forall j$$

Réponse polytomiale: interprétation (2)

• Avant d'exploiter cette sortie logicielle, calculons, pour une tranche d'âge donnée, les rapports de cotes des réactions des hommes vs les femmes directement à partir des données.

	Réaction				
Age	1	2	3	4	5
< 30	1.00	1.03	0.29	0.29	0.44
$30 - 39$	1.00	1.05	0.40	1.13	0.60
$40 - 49$	1.00	0.90	0.41	0.75	0.83
$50 - 64$	1.00	0.67	0.66	0.58	1.08
> 65	1.00	0.81	1.10	1.52	0.97

Ex : Le rapport des cotes de la `Reaction` '2' des femmes vs les hommes est estimé par $0.66/0.64 = 1.03$.

• L'absence d'interaction significative du 3ème ordre nous indique qu'il n'est pas nécessaire d'estimer ces RC séparément dans chacune des catégories d'`Age`.

• Les estimations fournies par le logiciel pour ces RC sont alors:

Réaction				
1	2	3	4	5
1.00	0.83	0.50	0.72	0.76

Ex Rapport des cotes (à un âge donné) de la réaction '3' des femmes vs les hommes: $\exp(-0.68699) = 0.50$: cela suggère que, à un âge donné, les femmes choisissent moins souvent que les hommes la réaction '3' au détriment de la '1'.

Réponse polytomiale: interprétation (3)

• L'examen des données révèlent effectivement que, quel que soit l'âge, l'infériorité de la popularité de 'Ni d'accord, ni en désaccord' par rapport à 'Tout à fait d'accord' est bien plus marquée chez les femmes que chez les hommes.

• Cependant, tous ces rapports de cotes ne sont pas significativement différents de 1: certains comptent 1 comme valeur plausible.

• La sortie logicielle révèle que la seule P-valeur $< 5\%$ réfère à la réponse $X = 3$.

• $\hat{\lambda}_{32}^{XZ} = -0.69$ (0.174) est, à un âge donné, le logarithme du rapport des cotes de $X = 3$ ('Ni d'accord, ni en désaccord') des femmes ($Z = 2$) versus les hommes ($Z = 1$).

• Les autres P-valeurs étant $> 5\%$, nous ne pouvons pas épingler de différence significative entre femmes et hommes pour les cotes correspondantes.

• Nous pouvons donc résumer nos estimations de RC (pour l'effet Sexe) par la table ci-contre où un '1' en italique indique que c'est une valeur plausible pour le RC, une autre valeur correspondant à un RC significativement différent de 1:

	Réaction			
1	2	3	4	5
1.00	*1*	0.50	*1*	*1*

Réponse polytomiale: interprétation (4)

Freq - Paramètres estimés (homo.sta)
Distribution : POISSON
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	Estimat.	Standard Erreur	Wald Stat.	p
Reaction*Age	1	34	1,77387	0,339508	27,299	0,000000
Reaction*Age	2	35	0,69642	0,356798	3,810	0,050954
Reaction*Age	3	36	0,54638	0,374510	2,128	0,144590
Reaction*Age	4	37	0,23281	0,401383	0,336	0,561909
Reaction*Age	5	38	0,00000			
Reaction*Age	6	39	0,96786	0,315964	9,383	0,002190
Reaction*Age	7	40	-0,09999	0,342529	0,085	0,770348
Reaction*Age	8	41	0,48305	0,301876	2,561	0,109561
Reaction*Age	9	42	-0,28022	0,366714	0,584	0,444793
Reaction*Age	10	43	0,00000			
Reaction*Age	11	44	0,82331	0,260063	10,022	0,001547
Reaction*Age	12	45	0,22983	0,245576	0,876	0,349343
Reaction*Age	13	46	-0,01925	0,266136	0,005	0,942347
Reaction*Age	14	47	-0,07225	0,266104	0,074	0,785993
Reaction*Age	15	48	0,00000			
Reaction*Age	16	49	0,49513	0,183339	7,293	0,006921
Reaction*Age	17	50	0,35670	0,153967	5,367	0,020520
Reaction*Age	18	51	0,15352	0,162392	0,894	0,344472
Reaction*Age	19	52	0,15488	0,161817	0,916	0,338488
Reaction*Age	20	53	0,00000			
Reaction*Age	21	54	0,00000			
Reaction*Age	22	55	0,00000			
Reaction*Age	23	56	0,00000			
Reaction*Age	24	57	0,00000			
Reaction*Age	25	58	0,00000			

Etiquettes des Colonnes (homo.sta)
Etiquettes des colonnes de la matrice X du modèle

Etiq.	Colonne	Variable	Niveau Variable	Variable	Niveau Variable
Reaction*Age	34	Reaction	5	Age	>65
Reaction*Age	35	Reaction	5	Age	50-64
Reaction*Age	36	Reaction	5	Age	40-49
Reaction*Age	37	Reaction	5	Age	30-39
Reaction*Age	38	Reaction	5	Age	<30
Reaction*Age	39	Reaction	4	Age	>65
Reaction*Age	40	Reaction	4	Age	50-64
Reaction*Age	41	Reaction	4	Age	40-49
Reaction*Age	42	Reaction	4	Age	30-39
Reaction*Age	43	Reaction	4	Age	<30
Reaction*Age	44	Reaction	3	Age	>65
Reaction*Age	45	Reaction	3	Age	50-64
Reaction*Age	46	Reaction	3	Age	40-49
Reaction*Age	47	Reaction	3	Age	30-39
Reaction*Age	48	Reaction	3	Age	<30
Reaction*Age	49	Reaction	2	Age	>65
Reaction*Age	50	Reaction	2	Age	50-64
Reaction*Age	51	Reaction	2	Age	40-49
Reaction*Age	52	Reaction	2	Age	30-39
Reaction*Age	53	Reaction	2	Age	<30
Reaction*Age	54	Reaction	1	Age	>65
Reaction*Age	55	Reaction	1	Age	50-64
Reaction*Age	56	Reaction	1	Age	40-49
Reaction*Age	57	Reaction	1	Age	30-39
Reaction*Age	58	Reaction	1	Age	<30

Réponse polytomiale: interprétation (5)

• La catégorie de référence pour `Réaction` est 1 ($X = 1$: 'Tout à fait d'accord') et celle pour `Age` est '< 30' ($Y = 1$).

• Puisque $\lambda_{ijk}^{XYZ} = 0$, les estimations réfèrent à

$$\lambda_{ik}^{XZ} = \log \frac{\text{Cote}(X = i|Y = j, Z = k)}{\text{Cote}(X = i|Y = 1, Z = k)} \quad \forall k$$

• L'absence d'interaction significative du 3ème ordre nous indique qu'il n'est pas nécessaire d'estimer ces RC séparément pour les hommes et les femmes.

• Comme précédemment, tous ces (logarithmes de) rapports de cotes ne sont pas significativement différents de 1 (0).

• De la sortie logicielle, on constate que les P-valeurs $< 5\%$ sont presque systématiquement associées à la catégorie des > 65 ans.

Réponse polytomiale: interprétation (6)

• Comme précédemment, un '1' en italique dans la table ci-contre indique que c'est une valeur plausible pour le RC, une autre valeur correspondant à un RC significativement différent de 1:

	Réaction				
Age	1	2	3	4	5
< 30	1.00	1.00	1.00	1.00	1.00
$30 - 39$	1.00	*1*	*1*	*1*	*1*
$40 - 49$	1.00	*1*	*1*	*1*	*1*
$50 - 64$	1.00	1.43	*1*	*1*	*1*
> 65	1.00	1.64	2.28	2.63	5.89

Ex Rapport des cotes (pour un sexe donné) de la réaction '5' des > 65 vs les < 30 ans: $\exp(1.77387) = 5.89$: cela suggère que, pour un sexe donné, les personnes de plus 65 ans choisissent (significativement) plus souvent que les mois de 30 ans la réaction '5' au détriment de la '1'.

• Ce contraste entre les > 65 ans et les < 30 ans est également significatif pour les autres niveaux de réponses ; il est d'autant plus marqué que le désaccord avec l'affirmation est prononcé.

• Un autre contraste significatif concerne la réaction '2' qui est plus souvent choisie au détriment de la réaction '1' par le 50-64 ans que par les < 30 ans.

ANNEXES

A.1 La distribution multinomiale *

• Pour rappel, $Y \sim \text{Bin}(n, \pi)$ lorsque Y comptabilise le nombre de succès en n expériences indépendantes, chacune ayant deux résultats possibles, le succès (avec probabilité π) ou l'échec (avec probabilité $1 - \pi$).

• L'espace d'échantillonnage est $\mathcal{E} = \{0, 1, \ldots, n\}$ et

$$\text{P}(Y = y) = \frac{n!}{y!\ (n-y)!}\ \pi^y (1-\pi)^{n-y} \quad \text{avec} \quad y \in \mathcal{E}$$

• Lorsque chaque expérience peut donner lieu à $K > 2$ résultats (mutuellement exclusifs) possibles avec probabilités respectives $\pi_1, \ldots, \pi_K$ et que Y_k compte le nombre de fois que le $k$ème résultat a été observés en n expériences indépendantes, alors

$$\text{P}(Y_1 = y_1, \ldots, Y_K = y_K | Y_+ = y_+) = \frac{n!}{y_1! \ldots y_K!}\ \pi_1^{y_1} \times \ldots \times \pi_K^{y_K}$$

avec $y_+ = y_1 + \ldots + y_K = n$, $y_k \in \{0, 1, \ldots, n\}$ et $\pi_1 + \ldots + \pi_K = 1$.

• Notation: $(Y_1, \ldots, Y_K) \sim \text{Mult}(n; \pi_1, \ldots, \pi_K)$.

A.2 Table de contingence et distribution multinomiale *

• Dans un des exemples précédents, $n = 1292$ personnes ont été questionnées et réparties en fonction de leur sexe et de leur (in)activité professionnelle.

• Chaque personne appartient ainsi à une et une seule des 4 catégories induites.

• La probabilité d'appartenir à la catégorie (i, j) vaut π_{ij}, la proportion d'individus dans la population avec ce statut (cf. plan de sondage...).

• Dans la table de contingence 2x2 de l'exemple, nous avons donc

$$(N_{11}, N_{21}, N_{12}, N_{22}) \sim \text{Mult}(n; \pi_{11}, \pi_{21}, \pi_{12}, \pi_{22})$$

A.3 Lien entre la Poisson et la multinomiale *

• Rappel: on dit que Y suit une distribution de Poisson de moyenne μ lorsque $\mathcal{E} = \{0, 1, 2, \ldots\}$ et $\mathrm{P}(Y = y) = \mathrm{e}^{-\mu}\,\dfrac{\mu^y}{y!}$. Notation: $Y \sim \mathrm{Pois}(\mu)$.

• On peut démontrer que si $Y_1 \sim \mathrm{Pois}(\mu_1), \ldots, Y_K \sim \mathrm{Pois}(\mu_K)$ sous la contrainte $Y_1 + \ldots + Y_K = n$, alors $(Y_1, \ldots, Y_K | Y_+ = n) \sim \mathrm{Mult}(n; \pi_1 = \dfrac{\mu_1}{n}, \ldots, \pi_K = \dfrac{\mu_K}{n})$.

• Par conséquent, si nous disposons d'une technique et/ou d'un logiciel permettant d'évaluer la distribution a posteriori des moyennes $\mu_1, \ldots, \mu_K$ de distributions de Poisson, on pourra facilement en déduire la distribution a posteriori des probabilités d'appartenance $\pi_1, \ldots, \pi_K$ (avec $\pi_k = \mu_k/n$) aux catégories $1, \ldots, K$.

Les données $y_1, \ldots, y_k$ seront simplement les fréquences associées à ces catégories.

• Dans notre exemple: le rôle des Y_k est joué par les N_{ij}. La distribution multinomiale décrite pour les féquences de la table de contingence peut être vue comme résultant de l'hypothèse

$$N_{ij} \sim \mathrm{Pois}(\mu_{ij}) \quad \text{avec} \quad \mu_{ij} = n \times \pi_{ij} \quad \text{et} \quad N_{11} + \ldots + N_{22} = n$$