Estimation d'une différence de risque standardisée à partir des sorties d'une régression logistique

Denis Talbot

Département de médecine sociale et préventive, Université Laval
Unité santé des populations et pratiques optimales en santé, CHU de Québec - Université Laval

Mai 2017

Plan

- Mise en situation
- Calcul d'une différence de risque standardisée avec SAS
  - Par modélisation de l'issue (*g-formula*)
  - Par modélisation de l'exposition (pondération)
- Conclusion

Mise en situation

Question de recherche : Quelle est l'association entre la consommation de lait au cours de la vie et le risque de fracture de la hanche ?

Le National Health and Nutrition Examination Survey (NHANES)

- Vaste étude annuelle aux États-Unis sur la santé et l'alimentation, combinant questionnaires et examens médicaux
- Nous considérons les données de 5935 participants récoltées entre janvier 2007 et décembre 2008
- Certaines modificaitons aux données ont été apportées pour simplifier l'illustration

Variables d'intérêt principal

Exposition (X) : Consommation de lait au cours de la vie - $X = 1$ s'il y a eu consommation régulière à au moins un moment, $X = 0$ sinon.

Issue (Y) : Occurence d'une fracture de la hanche - $Y = 1$ si le sujet à eu une fracture de la hanche et $Y = 0$ sinon.

Statistiques descriptives

	Non consom. (n = 1467)	Consom. (n = 4468)
Hommes, n (%)	635 (43,3%)	2275 (50,9%)
Race/Ethnicité, n (%)		
- Mexicains	280 (19,1%)	753 (16,9%)
- Autres hispaniques	166 (11,3%)	500 (11,2%)
- Blancs	599 (40,8%)	2162 (48,4%)
- Noirs	340 (23,2%)	887 (19,9%)
- Autres	82 (5,6%)	166 (3,7%)
Prednisone ou cortisone, n (%)	67 (4,6%)	213 (4,8%)
Fracture de la mère, n (%)	64 (4,4%)	251 (5,6%)
Fracture du père, n (%)	39 (2,7%)	91 (2,0%)
Alcool (O/N), n (%)	1002 (68,3%)	3144 (70,4%)
Âge en années, moy (ET)	50,7 (17,6)	50,8 (18,1)
Alcool nb/j (?), moy (ET)	2,7 (2,7)	2,6 (2,9)
IMC en kg/m^2, moy (ET)	29,0 (6,6)	29,0 (6,7)

Association

```
PROC FREQ DATA = osteo2;
    TABLE cons*OSQ010A / CHISQ RISKDIFF;
RUN;
```

Frequency Percent Row Pct Col Pct	Table of cons by OSQ010A		
	OSQ010A		
cons	0	1	Total
0	1440 24.26 98.16 24.56	27 0.45 1.84 37.50	1467 24.72
1	4423 74.52 98.99 75.44	45 0.76 1.01 62.50	4468 75.28
Total	5863 98.79	72 1.21	5935 100.00

Fractures non exposés = 27/1467 = 1,84%
Fractures exposés = 45/4468 = 1,01%
Différence de proportion = -0,83%
Rapport de proporition = 0,54

Pourquoi favoriser l'échelle additive ?

Permet plus directement de comparer l'impact sur la santé publique de différentes interventions.

Deux expositions peuvent avoir un rapport de proportions différents, mais être associées à un même nombre de cas. Exemple :

Proportion non exposés = 1,84%, proportion exposés = 1,01% ; DP = -0,83%, RP = 0,54.

Proportion non exposés = 1,34%, proportion exposés = 0,51% ; DP = -0,83%, RP = 0,38.

Différence standardisée

Les groupes comparés diffèrent dans leurs caractéristiques.
La différence dans le risque de fracture pourrait être attribuable à ces différences.

La **standardisation** permet de rendre comparables deux groupes en ce qui a trait à certaines caractéristiques.

La standardisation est à la base de plusieurs méthodes d'inférence causale à partir de données observationnelles.

Comment fonctionne la standardisation ?

Intuitivement, pour chaque groupe

- On divise la population en strates en fonction des valeurs des variables de standardisation
- Chaque strate reçoit un poids correspondant à son importance dans la population
- On effectue une moyenne pondérée sur les strates de la caractéristique d'intérêt

Mathématiquement la moyenne standardisée pour L de Y dans le groupe $X = x$ est $\sum_l \mathbb{E}[Y|X = x, L]P(L = l)$

Approche classique pour calculer une association ajustée

Ajuster un modèle de régression logistique de l'occurence de fracture de la hanche (Y) en fonction de la consommation de lait (X) conditionnelle aux variables potentiellement confondantes (L).

$$logit[P(Y = 1|X, L)] = \beta_0 + \beta_1 X + \beta_2 L$$

```
PROC GENMOD DATA = osteo2 DESCENDING;
    CLASS cons (REF = '0');
    MODEL OSQ010A = cons RIAGENDR RIDAGEYR OSQ130 OSQ170 OSQ200
    ALQ101 ALQ130 BMXBMI mex_am aut_hisp blanc_nh noir_nh
        / DIST = binomial LINK = logit;
    LSMEANS cons / CL ILINK;
RUN;
```

Problèmes avec l'approche classique

- L'exponentiel du paramètre s'interprète comme un rapport de cote
- Les moyennes des moindres carrés (`LSMEANS`) donnent des moyennes pour un sujet « moyen » ($L = \bar{L}$)
- Il n'y a pas de moyen simple direct d'obtenir une différence de risque standardisée

Standardisation par modélisation de l'issue

Les sorties de la régression logistique permettent de calculer la différence de proportion standardisée.

Pour chacune des strates des catégories de L (l), on peut estimer

- La proportion de fracture en absence de consommation de lait $\hat{P}(Y = 1|X = 0, L = l) = expit\{\hat{\beta}_0 + \hat{\beta}_2 l\}$
- La proportion de fracture en présence de consommation de lait $\hat{P}(Y = 1|X = 1, L = l) = expit\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 l\}$
- La différence de proportion $expit\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 l\} - expit\{\hat{\beta}_0 + \hat{\beta}_2 l\}$

Il ne reste qu'à effectuer une moyenne pondérée en fonction de l'importance des strates.

Standardisation par modélisation de l'issue avec SAS

1. Créer une base de données augmentée contenant trois copies de chaque observation
   - 1.1 Une copie originale.
   - 1.2 Une copie où tous les sujets sont non-exposés et l'issue est manquante ($X = 0, Y = .$)
   - 1.3 Une copie où tous les sujets sont exposés et l'issue est manquante ($X = 1, Y = .$)
2. Ajuster le modèle de régression logistique sur la base de données augmentée.
3. Calculer et comparer la proportion de fracture sur les copies « non-exposés » et « exposés »

```
DATA osteo2_augm;
    SET osteo2;
    copie = 0;
    OUTPUT;
    copie = 1;
    cons = 0;
    OSQ010A = .;
    OUTPUT;
    copie = 1;
    cons = 1;
    OSQ010A = .;
    OUTPUT;
RUN;
```

```
PROC GENMOD DATA = osteo2_augm DESCENDING;
    CLASS cons (REF = '0');
    MODEL OSQ010A = cons RIAGENDR RIDAGEYR OSQ130 OSQ170 OSQ200
    ALQ101 ALQ130 BMXBMI mex_am aut_hisp blanc_nh noir_nh
        / DIST = binomial LINK = logit;
    OUTPUT OUT = sortie_pred P = predit;
RUN;
```

```
PROC MEANS DATA = sortie_pred;
    WHERE copie = 1;
    CLASS cons;
    VAR predit;
RUN;
```

PROC SURVEYSELECT pour obtenir les échantillons bootstraps pour construire un IC.

```
PROC SURVEYSELECT DATA = osteo2
    METHOD = URS
    SAMPSIZE = 5935
    SEED = 58109571
    OUTHITS
    OUT = osteo2_boot
    REP = 2000;
RUN;
```

Le reste de la programmation demande des ajustements relativement mineurs.

Macro pour obtenir la différence standardisée

```
%INCLUDE "C:\Users\detal9\Dropbox\Travail\Présentations\Club SAS 2017\counterfactual.sas";

%counterfactual(data = osteo2, yvar = OSQ010A, trt = cons,
    xvar = RIAGENDR RIDAGEYR OSQ130 OSQ170 OSQ200
    ALQ101 ALQ130 BMXBMI mex_am aut_hisp blanc_nh noir_nh,
    link = logit, alpha = 0.05);
```

Effectue un calcul théorique de la variance (méthode du delta).

Résultats

	group1								
	ESTIMATE	WALD_L	WALD_U	LOG_L	LOG_U	LGT_L	LGT_U	WLSN_L	WLSN_U
ROW1	0.0099765	0.0070855	0.0128676	0.0074667	0.0133301	0.0074638	0.0133239	0.0074715	0.0133103

	group0								
	ESTIMATE	WALD_L	WALD_U	LOG_L	LOG_U	LGT_L	LGT_U	WLSN_L	WLSN_U
ROW1	0.0189314	0.0118737	0.0259892	0.01304	0.0274847	0.0130244	0.027443	0.0130534	0.027383

dWald		
-0.008955	-0.016599	-0.001311

Proportion de fractures standardisée chez les non exposés = 1,89%
Proportion de fractures standardisée chez exposés = 1,00%
Différence de proportion standardisée = -0,90%
IC à 95% bootstrap : -1,69% à -0,18% ;
IC à 95% Wald : -1,66% à -0,13%

Standardisation par modélisation de l'exposition

Chaque sujet reçoit un poids correspondant à l'inverse de la probabilité de son niveau d'exposition conditionnellement à ses variables potentiellement confondantes.

$$w_i = \frac{1}{P(X = x_i | L = l_i)}$$

Ces poids créent une pseudo-population où l'exposition n'est plus associée aux variables L...

Standardisation par modélisation de l'exposition

Chaque sujet reçoit un poids correspondant à l'inverse de la probabilité de son niveau d'exposition conditionnellement à ses variables potentiellement confondantes.

$$w_i = \frac{1}{P(X = x_i | L = l_i)}$$

Ces poids créent une pseudo-population où l'exposition n'est plus associée aux variables L... **la distribution de L est la même que dans la population d'origine**.

Standardisation par modélisation de l'exposition avec SAS

```
PROC LOGISTIC DATA = osteo2 DESCENDING;
    MODEL cons = RIAGENDR RIDAGEYR OSQ130 OSQ170 OSQ200
    ALQ101 ALQ130 BMXBMI
    mex_am aut_hisp blanc_nh noir_nh;
    OUTPUT OUT = sortie P = pred;
RUN;
```

```
DATA poids;
    SET sortie;
    IF cons = 1 THEN w = 1/pred;
    IF cons = 0 THEN w = 1/(1 - pred);
RUN;
```

```
PROC GENMOD DATA = poids DESCENDING;
    CLASS SEQN cons(REF = '0');
    MODEL OSQ010A = cons / DIST = binomial LINK = ID;
    LSMEANS cons / DIFF CL;
    REPEATED SUBJECT = SEQN;
    WEIGHT w;
RUN;
```

Résultats

Proportion de fractures standardisée chez les non exposés = 1,83%
Proportion de fractures standardisée chez exposés = 1,00%
Différence de proportion standardisée = -0,83%
IC à 95% : -1,59% à -0,08%.

Conclusion

- L'ajustement peut être nécessaire pour bien établir l'effet de l'exposition lorsque les groupes diffèrent.

Conclusion

- L'ajustement peut être nécessaire pour bien établir l'effet de l'exposition lorsque les groupes diffèrent.
- L'approche classique d'ajustement donne un rapport de cote qui peut être difficile de traduire en termes d'impact sur la santé publique.

Conclusion

- L'ajustement peut être nécessaire pour bien établir l'effet de l'exposition lorsque les groupes diffèrent.
- L'approche classique d'ajustement donne un rapport de cote qui peut être difficile de traduire en termes d'impact sur la santé publique.
- La différence de risque standardisée est une solution alternative qui peut répondre plus clairement aux questions d'intérêt.

Conclusion

- L'ajustement peut être nécessaire pour bien établir l'effet de l'exposition lorsque les groupes diffèrent.
- L'approche classique d'ajustement donne un rapport de cote qui peut être difficile de traduire en termes d'impact sur la santé publique.
- La différence de risque standardisée est une solution alternative qui peut répondre plus clairement aux questions d'intérêt.
- Cette différence peut être assez facilement calculée à l'aide de SAS.

Conclusion

- L'ajustement peut être nécessaire pour bien établir l'effet de l'exposition lorsque les groupes diffèrent.
- L'approche classique d'ajustement donne un rapport de cote qui peut être difficile de traduire en termes d'impact sur la santé publique.
- La différence de risque standardisée est une solution alternative qui peut répondre plus clairement aux questions d'intérêt.
- Cette différence peut être assez facilement calculée à l'aide de SAS.
- Les résultats peuvent varier en fonction de l'approche choisie pour standardisé en raison des hypothèses de modélisation.

Références

- Zou GY. Assessment of risks by predicting counterfactuals. Statistics in medicine. 2009 ;28(30) :3761-81.
- Muller CJ, MacLehose RF. Estimating predicted probabilities from logistic regression : different methods correspond to different target populations. International journal of epidemiology. 2014 ;43(3) :962-70.
- Hernán MA, Robins JM. Causal Inference. 2017. Boca Raton : Chapman & Hall/CRC, forthcoming
  - Chapitre 12 - pondération
  - Chapitre 13 - formule-g

Questions ?

Note : Tous les fichiers sont disponibles sur mon site Internet
`https://sites.google.com/site/denistalbotfmed/autres`