Transmission desequilibrium test et régression logistique conditionnelle.

Rules and Tools, INRA, Toulouse

Hervé Perdry

Université Paris-Sud UMR-S 669 & INSERM U669

16 février 2012

TDT et régression logistique conditionnelle

Demandez le programme

- Présentation du TDT ; motivation, test du TDT, estimation des risques relatifs.
- Présentation de la régression logistique conditionnelle ;
- Utilisation de la régression logistique conditionnelle avec « pseudo-contrôles » pour réaliser le TDT.

- Tout ceci pour le cas simple d'un marqueur di-allélique ! Ne seront pas évoqués, en particulier : tests pour un effet parent d'origine, tests haplotypiques, tests d'interaction, etc.

Nous essayons de montrer que les techniques « savantes » mises en place à partir du simple TDT en sont des prolongements naturels, ce qui peut aider à leur compréhension et à leur interprétation.

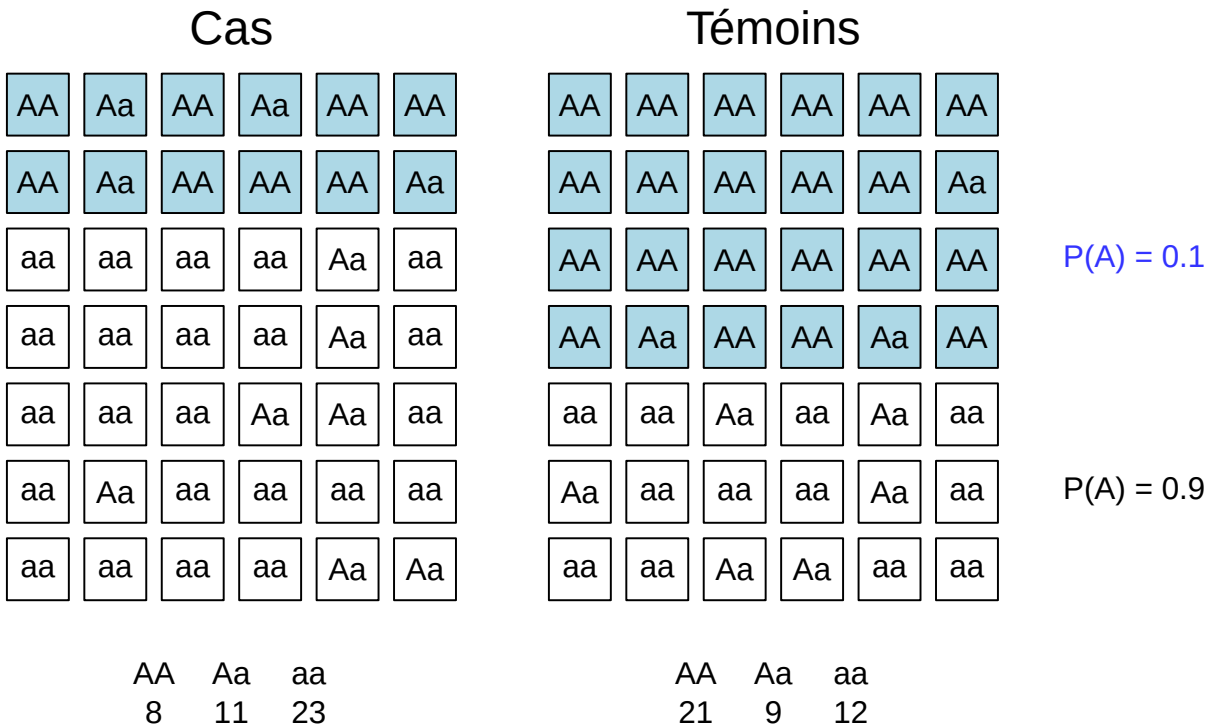
TDT et régression logistique conditionnelle

1. Le Transmission Disequilibrium Test (TDT)
2. Risques relatifs
3. La régression logistique conditionnelle (RLC)
4. Le TDT par la RLC

Motivation

Étude cas/témoins et stratification de population

Dans une étude cas/témoins, si le recrutement a lieu dans une population stratifiée, il peut arriver ça :

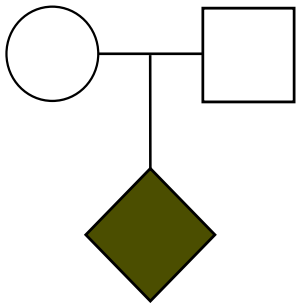


☞ la sous-population est une variable de confusion

Également : facteur environnemental impliqué dans la maladie et corrélation gène-environnement (e.g. gradient nord/sud).

Le Transmission Disequilibrium Test

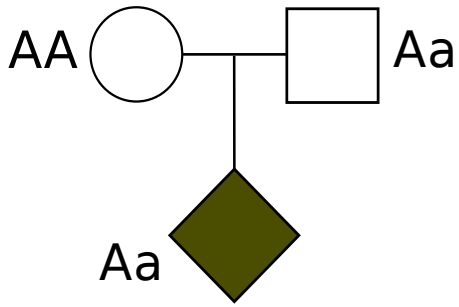
Le TDT traite des familles trio :



Deux parents et un enfant atteint.

Le Transmission Disequilibrium Test

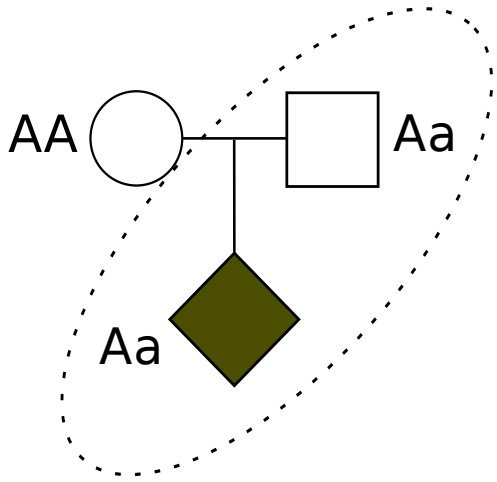
Le TDT traite des familles trio :



Deux parents et un enfant atteint. Leurs génotypes à un marqueur di-allélique, d'allèles A et a.

Le Transmission Disequilibrium Test

Le TDT traite des familles trio :



Deux parents et un enfant atteint. Leurs génotypes à un marqueur di-allélique, d'allèles A et a.

L'événement informatif est la transmission de A ou a par un parent hétérozygote.

Sous l'hypothèse nulle, A est transmis aussi souvent que a.

Le Transmission Disequilibrium Test

- Les parents homozygotes ne fournissent aucune information ;
- on compte le nombre de transmissions des allèles A et a par un parent hétérozygote.

génotypes			transmis	
père	mère	enfant	A	a
AA	Aa	AA	1	0
Aa	AA	Aa	0	1
AA	AA	AA	0	0
Aa	Aa	Aa	1	1
Aa	Aa	AA	2	0
⋮				
		Total	n_A	n_a

Le Transmission Disequilibrium Test

On compare ensuite le nombre de transmissions de A et a, respectivement n_A et n_a, au nombre attendu sous H_0 : « probabilité pour un parent Aa de transmettre A $= \frac{1}{2}$ ».

☞ test du χ^2

	A	a	Total
Observés	n_A	n_a	$n = n_A + n_a$
Attendus	$\frac{1}{2}n$	$\frac{1}{2}n$	n

$$\chi^2(1) = \frac{\left(n_A - \frac{1}{2}n\right)^2}{\frac{1}{2}n} + \frac{\left(n_a - \frac{1}{2}n\right)^2}{\frac{1}{2}n} = \frac{(n_A - n_a)^2}{n_A + n_a}.$$

Le Transmission Disequilibrium Test

Précisons l'hypothèse nulle, sous un modèle simple (simpliste ?).

On considère une population (stratifiée). Un unique locus maladie di-allélique D/d. On note π_D (resp. π_d) les probabilités pour un hétérozygote Dd de transmettre D (resp. d) à un enfant atteint : on a $\pi_D > \pi_d$ (par exemple).

Autres paramètres :

- $\mathbb{P}(D|A)$, $\mathbb{P}(D|a)$, $\mathbb{P}(d|A) = 1 - \mathbb{P}(D|A)$, $\mathbb{P}(d|a) = 1 - \mathbb{P}(D|a)$ caractérisent le déséquilibre gamétique dans la population ; pas de déséquilibre si $\mathbb{P}(D|A) = \mathbb{P}(D|a)$.
- le taux de recombinaison θ entre le marqueur et le locus maladie ; pas de liaison si $\theta = \frac{1}{2}$.

Le Transmission Disequilibrium Test

On considère un parent hétérozygote Aa. S'il est homozygote DD (resp. dd), tous les gamètes portent l'allèle D (resp. d) et il n'y a pas de distortion du taux de transmission : $\mathbb{P}(\text{transmet } A | DD \text{ ou } dd) = \frac{1}{2}$.

Il est hétérozygote Dd avec probabilité

$$\begin{aligned} &\mathbb{P}(D|A)\mathbb{P}(d|a) && \text{phase DA / da} \\ +&\mathbb{P}(d|A)\mathbb{P}(D|a) && \text{phase dA / Aa .} \end{aligned}$$

Si la phase est DA / da, les gamètes sont DA avec probabilité $\frac{1}{2}(1-\theta)$, Da avec probabilité $\frac{1}{2}\theta$, etc.

Pour finir :

$$\begin{aligned} \mathbb{P}(\text{transmet } A) = &\frac{1}{2}\mathbb{P}(D|A)\mathbb{P}(d|a)((1-\theta)\pi_D + \theta\pi_d) \\ &+ \frac{1}{2}\mathbb{P}(d|A)\mathbb{P}(D|a)(\theta\pi_D + (1-\theta)\pi_d) \\ &+ \frac{1}{2}(1 - \mathbb{P}(D|a)\mathbb{P}(d|a) - \mathbb{P}(d|A)\mathbb{P}(D|a)) \end{aligned}$$

Le Transmission Disequilibrium Test

On voit que dès que $\theta = \frac{1}{2}$ ou $\mathbb{P}(D|A) = \mathbb{P}(D|a)$, on a $\mathbb{P}(\text{transmet } A) = \frac{1}{2}$.

☞ $H_0 =$ « $\theta = \frac{1}{2}$ ou $\mathbb{P}(D|A) = \mathbb{P}(d|a)$ »
$=$ « pas de liaison ou pas de déséquilibre ».

Le TDT est un test d'association et de liaison simultanée ; ou encore, un test de *déséquilibre de liaison* (par opposition au déséquilibre gamétique) entre le marqueur et le locus maladie.

TDT et régression logistique conditionnelle

Risques relatifs et odds ratio

Étude cas/témoins

La p-valeur n'apprend rien sur l'effet ; elle dépend à la fois de la taille de l'échantillon et de la taille de l'effet.

☞ Intérêt pour les risques relatifs.

$$f_0 = \mathbb{P}(\text{cas}|AA) = \frac{\mathbb{P}(AA|\text{cas})\mathbb{P}(\text{cas})}{\mathbb{P}(AA)} \simeq \frac{\mathbb{P}(AA|\text{cas})\mathbb{P}(\text{cas})}{\mathbb{P}(AA|\text{témoin})},$$

$$f_1 = \mathbb{P}(\text{cas}|Aa) = \frac{\mathbb{P}(Aa|\text{cas})\mathbb{P}(\text{cas})}{\mathbb{P}(Aa)} \simeq \frac{\mathbb{P}(Aa|\text{cas})\mathbb{P}(\text{cas})}{\mathbb{P}(Aa|\text{témoin})},$$

$$\psi_1 = \frac{\mathbb{P}(\text{cas}|Aa)}{\mathbb{P}(\text{cas}|AA)} \simeq \frac{\mathbb{P}(Aa|\text{cas})\mathbb{P}(AA|\text{témoin})}{\mathbb{P}(Aa|\text{témoin})\mathbb{P}(AA|\text{cas})}.$$

Risques relatifs et odds ratio

Étude cas/témoins

	AA	Aa	aa
cas	8	11	23
témoins	21	9	12

$$\psi_1 = \frac{\mathbb{P}(\text{cas}|Aa)}{\mathbb{P}(\text{cas}|AA)} \simeq \frac{11 \cdot 21}{9 \cdot 8} \simeq 3.21, \ \psi_2 = \frac{\mathbb{P}(\text{cas}|aa)}{\mathbb{P}(\text{cas}|AA)} \simeq \frac{23 \cdot 21}{12 \cdot 8} \simeq 5.03.$$

Cette approximation des risques relatifs est connue sous le nom d'odds ratio (OR), ce qui correspond à l'écriture

$$\frac{\mathbb{P}(Aa|\text{cas})/\mathbb{P}(Aa|\text{témoin})}{\mathbb{P}(AA|\text{cas})/\mathbb{P}(AA|\text{témoin})} = \frac{\mathbb{P}(\text{cas}|Aa)/\mathbb{P}(\text{témoin}|Aa)}{\mathbb{P}(\text{cas}|AA)/\mathbb{P}(\text{témoin}|AA)}.$$

Risques relatifs pour le TDT

On veut pouvoir travailler avec une population stratifiée ; attention à ne pas écrire un modèle incluant l'équilibre de Hardy-Weinberg.

Une solution (Schaid & Sommer, AJHG, 1993) : la vraisemblance conditionnelle aux génotypes parentaux (CPG : *conditional on parental genotypes)*. On ne s'intéresse pas à la probabilité des appariements *(mating types)*, qui n'est pas informative.

Comme précédement, on note les risques $f_0 = \mathbb{P}(\text{cas}|AA)$, $f_1 = \mathbb{P}(\text{cas}|Aa)$, $f_2 = \mathbb{P}(\text{cas}|aa)$, et les risques relatifs $\psi_1 = f_1/f_0$ et $\psi_2 = f_2/f_0$.

La vraisemblance de chaque trio est la probabilité du génotype observé chez le cas, conditionnellement aux génotypes parentaux. Elle s'exprime uniquement avec ψ_1 et ψ_2.

Risques relatifs pour le TDT

Par exemple, la vraisemblance d'un trio avec parents AA et Aa et enfant AA est (on n'écrit pas le conditionnement aux parents, implicite pour toutes les probas qui suivent) :

$$\begin{aligned}\mathbb{P}(\text{enfant AA}|\text{enfant atteint}) &= \frac{\mathbb{P}(\text{enfant atteint}|\text{enfant AA})\mathbb{P}(\text{enfant AA})}{\mathbb{P}(\text{enfant atteint})} \\ &= \frac{f_0 \times \frac{1}{2}}{f_0 \times \frac{1}{2} + f_1 \times \frac{1}{2}} \\ &= \frac{1}{1+\psi_1}\end{aligned}$$

et pour un enfant Aa :

$$\mathbb{P}(\text{enfant Aa}|\text{enfant atteint}) = \frac{\psi_1}{1+\psi_1}.$$

Risques relatifs pour le TDT

La CPG : tous les appariements possibles

Appariement	Génotype du cas	$\ell(\psi_1, \psi_2)$
AA × AA	AA	1
AA × Aa	AA	$\frac{1}{1+\psi_1}$
	Aa	$\frac{\psi_1}{1+\psi_1}$
AA × aa	Aa	1
Aa × Aa	AA	$\frac{1}{1+2\psi_1+\psi_2}$
	Aa	$\frac{2\psi_1}{1+2\psi_1+\psi_2}$
	aa	$\frac{\psi_2}{1+2\psi_1+\psi_2}$
Aa × aa	Aa	$\frac{\psi_1}{\psi_1+\psi_2}$
	aa	$\frac{\psi_2}{\psi_1+\psi_2}$
aa × aa	aa	1

Risques relatifs pour le TDT

On écrit la vraisemblance de l'échantillon de trios comme le produit des vraisemblances de chacun des trios, données par la table précédente.

- ☞ estimation de ψ_1 et ψ_2 par maximum de vraisemblance ;
- ☞ possibilité de tester modèles « additif » $(\psi_2 = \psi_1^2)$, dominant $(\psi_1 = \psi_2)$, récessif $(\psi_1 = 1)$...

Risques relatifs pour le TDT

On écrit la vraisemblance de l'échantillon de trios comme le produit des vraisemblances de chacun des trios, données par la table précédente.

- ☞ estimation de ψ_1 et ψ_2 par maximum de vraisemblance ;
- ☞ possibilité de tester modèles « additif » $(\psi_2 = \psi_1^2)$, dominant $(\psi_1 = \psi_2)$, récessif $(\psi_1 = 1)$...

Remarque : Dans le cas du modèle additif, on vérifie que $\widehat{\psi}_1 = n_a/n_A$ (avec les notations introduites pour le TDT). Le test associé à ce modèle est équivalent au TDT, qui n'a pourtant pas été conçu comme un test du modèle additif.

TDT et régression logistique conditionnelle

La régression logistique

Observations $(Y, X) = (Y_i, X_i)_{i=1,\dots,n}$, où $Y_1 \in \{0,1\}$ (eg sain et malade) et $X_i \in \mathbb{R}^d$ variable (supposée) prédictrice, eg le génotype en un SNP, qu'on recode par une variable $X_i = 0,1,2$ pour un modèle « additif », ou bien $X_i = (X_{i1}, X_{i2})$ pour un modèle « génotypique », comme ci-dessous :

génotype	X_i	X_{i1}	X_{i2}
AA	0	0	0
Aa	1	1	0
aa	2	0	1

On suppose qu'on a entre Y_i et X_i une relation du type

$$\mathbb{P}(Y_i = 1 | X_i = x_i) = \pi_i = \text{logit}^{-1}(\alpha + \beta' x_i) = \frac{e^{\alpha+\beta' x_i}}{1 + e^{\alpha+\beta' x_i}}.$$

La régression logistique

La vraisemblance d'une observation $(Y_i = y_i, X_i = x_i)$ est

$$\ell_i(\alpha,\beta;y_i,x_i) = \begin{cases} \pi_i = \dfrac{e^{\alpha+\beta' x_i}}{1+e^{\alpha+\beta' x_i}} \text{ si } y_i = 1, \\ 1-\pi_i = \dfrac{1}{1+e^{\alpha+\beta' x_i}} \text{ si } y_i = 0, \end{cases}$$

ou

$$\ell_i(\alpha,\beta;y_i,x_i) = \frac{e^{(\alpha+\beta' x_i)y_i}}{1+e^{(\alpha+\beta' x_i)}}.$$

La vraisemblance de la totalité de l'échantillon s'obtient en prenant le produit des $\ell_i(\alpha,\beta)$:

$$\ell(\alpha,\beta;y,x) = \frac{\prod_i e^{(\alpha+\beta' x_i)y_i}}{\prod_i \left(1+e^{\alpha+\beta' x_i}\right)} = \frac{e^{\alpha\sum y_i}\, e^{\sum \beta' x_i y_i}}{\prod_i \left(1+e^{\alpha+\beta' x_i}\right)}.$$

☞ Estimation par maximum de vraisemblance, test du score, etc.

La régression logistique

Exemple : données cas/témoins, un SNP

	AA	Aa	aa
cas	8	11	23
témoins	21	9	12

Considérons le modèle génotypique :

$$\text{logit}\,\mathbb{P}(Y_1 = 1 | x_{i1}, x_{i2}) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

On note $\pi_{AA} = \frac{e^{\alpha}}{1+e^{\alpha}}$, $\pi_{Aa} = \frac{e^{\alpha+\beta_1}}{1+e^{\alpha+\beta_1}}$ et $\pi_{aa} = \frac{e^{\alpha+\beta_2}}{1+e^{\alpha+\beta_2}}$.

Il y a une correspondance 1 à 1 entre $(\alpha, \beta_1, \beta_2)$ et $(\pi_{AA}, \pi_{Aa}, \pi_{aa})$.

La régression logistique

Exemple : données cas/témoins, un SNP

	AA	Aa	aa
cas	8	11	23
témoins	21	9	12

La vraisemblance est

$$(\pi_{AA})^8(\pi_{Aa})^{11}(\pi_{aa})^{23}(1-\pi_{AA})^{21}(1-\pi_{Aa})^9(1-\pi_{aa})^{12},$$

et il est facile de vérifier qu'elle est maximale pour

$$\pi_{AA} = \frac{8}{8+21},\ \pi_{Aa} = \frac{11}{11+9},\ \pi_{aa} = \frac{23}{23+12}.$$

La régression logistique

Exemple : données cas/témoins, un SNP

On sait qu'en pareil cas, les valeurs des $\pi_{AA}, \pi_{Aa}, \pi_{aa}$ n'ont pas d'interprétation propre, seul les rapports entre les odds $\pi_{AA}/(1-\pi_{AA})$, $\pi_{Aa}/(1-\pi_{Aa})$, et $\pi_{aa}/(1-\pi_{aa})$ ont un sens en tant qu'approximation des risques relatifs. En effet, par exemple :

$$\frac{\pi_{Aa}/(1-\pi_{Aa})}{\pi_{AA}/(1-\pi_{AA})} = \frac{\mathbb{P}(Aa|\text{cas})/\mathbb{P}(Aa|\text{témoins})}{\mathbb{P}(AA|\text{cas})/\mathbb{P}(AA|\text{témoins})} \simeq \frac{\mathbb{P}(\text{cas}|Aa)}{\mathbb{P}(\text{cas}|AA)}.$$

Si on revient aux notations en β, on retrouve le résultat bien connu

$$e^{\beta_1} = \frac{\pi_{Aa}/(1-\pi_{Aa})}{\pi_{AA}/(1-\pi_{AA})}, \quad e^{\beta_2} = \frac{\pi_{aa}/(1-\pi_{aa})}{\pi_{AA}/(1-\pi_{AA})}.$$

☞ Avec ce modèle, la régression logistique est identique à un traitement élémentaire « classique » de la table de contingence.

	AA	Aa	aa
cas	8	11	23
témoins	21	9	12

Pour conclure cet exemple :

- le modèle génotypique mène à un test à deux ddl qui est significatif, et les estimations $\widehat{\beta}_1 \simeq 1.17$ et $\widehat{\beta}_2 \simeq 1.62$ correspondent bien aux valeurs attendues $\log(3.21)$ et $\log(5.03)$.
- le modèle additif mène à un test à un ddl, également significatif, et à $\widehat{\beta} \simeq 0.79$.

La régression logistique

Le modèle sous-jacent suppose implicitement qu'on a recueilli n observations, puis observé les valeurs des Y_i.

Cependant dans une étude cas/témoins classique, on planifie le recueil de n_1 cas et n_0 témoins.

- ☞ dans ce cas, l'estimation de l'« intercept » α n'a plus ni sens ni intérêt.
- ☞ il peut être jugé naturel ou préférable de considérer la vraisemblance des observations *conditionnellement* au plan d'expérience !

La régression logistique conditionnelle

La vraisemblance conditionnelle :

$$\begin{aligned}\ell(\alpha,\beta;y,x) = \ell(\beta;y,x) &= \mathbb{P}(Y=y|X=x,\textstyle\sum_i y_i = n_1)\\ &= \frac{e^{\alpha\sum y_i}\, e^{\sum\beta' x_i y_i} \Big/ \prod_i \left(1+e^{\alpha+\beta' x_i}\right)}{\sum_{\nu\in S} e^{\alpha\sum \nu_i}\, e^{\sum\beta' x_i \nu_i} \Big/ \prod_i \left(1+e^{\alpha+\beta' x_i}\right)}\\ &= \frac{e^{\sum\beta' x_i y_i}}{\sum_{\nu\in V} e^{\sum\beta' x_i \nu_i}},\end{aligned}$$

où V est l'ensemble des vecteurs $\nu = (\nu_1,\ldots,\nu_n)$ avec $\nu_i = 0,1$ et $\sum_i \nu_i = n_1$.

Le dénominateur est la probabilité d'avoir observé n_1 cas parmi les n observations, on a sommé sur toutes les répartitions possibles pour les cas.

Le conditionnement fait disparaître l'intercept α.

La régression logistique conditionnelle

Exemple : un cas $y_1 = 1$ et deux témoins $y_2, y_3 = 0$, et $x_1, x_2, x_3 \in \mathbb{R}$. La vraisemblance pour la régression logistique est

$$\frac{e^{\alpha+\beta x_1}}{(1+e^{\alpha+\beta x_1})(1+e^{\alpha+\beta x_2})(1+e^{\alpha+\beta x_3})},$$

et pour la régression logistique conditionnelle

$$\frac{e^{\beta x_1}}{e^{\beta x_1}+e^{\beta x_2}+e^{\beta x_3}}.$$

La régression logistique conditionnelle

Exemple : deux cas $y_1, y_2 = 1$ et deux témoins $y_3, y_4 = 0$, et $x_1, x_2, x_3, x_4 \in \mathbb{R}$.
La vraisemblance pour la régression logistique est

$$\frac{e^{2\alpha+\beta(x_1+x_2)}}{(1+e^{\alpha+\beta x_1})(1+e^{\alpha+\beta x_2})(1+e^{\alpha+\beta x_3})(1+e^{\alpha+\beta x_4})},$$

et pour la régression logistique conditionnelle

$$\frac{e^{\beta(x_1+x_2)}}{e^{\beta(x_1+x_2)}+e^{\beta(x_1+x_3)}+e^{\beta(x_1+x_4)}+e^{\beta(x_2+x_3)}+e^{\beta(x_2+x_4)}+e^{\beta(x_3+x_4)}}.$$

La régression logistique conditionnelle

Les vraisemblances diffèrent : les estimations diffèrent ! Un petit jeu de données en exemple :

x	y	x	y	x	y	x	y
2.66	0	2.06	0	9.35	1	4.82	0
3.72	0	1.77	0	2.12	0	6.00	1
5.73	1	6.87	1	6.52	1	4.94	0
9.08	1	3.84	0	1.26	0	1.86	0
2.02	0	7.70	1	2.67	0	8.27	1
8.98	1	4.98	0	3.86	0	6.68	1
9.45	1	7.18	1	0.13	0	7.94	1
6.61	1	9.92	1	3.82	0	1.08	0
6.29	1	3.80	0	8.70	1	7.24	1
0.62	0	7.77	1	3.40	0	4.11	1

La régression logistique estime $\widehat{\alpha} = 0.299$, $\widehat{\beta} = 0.110$.
La régression logistique conditionnelle estime $\widehat{\beta} = 0.107$.

La régression logistique conditionnelle

Expérience en strates : on recueille les données (Y_{ki}, X_{ki}) dans des strates $S_1, \ldots, S_K$, de façon a avoir n_{k0} témoins et n_{k1} cas dans la strate k.

Dans chaque strate S_k on a un modèle

$$\text{logit}\,\mathbb{P}(Y_{ki} = 1 | X_{ki} = x_{ki}) = \alpha_k + \beta' x_{ki}.$$

Risque de base différent dans chaque strate ; eg facteurs environnementaux non observés mais constants dans chaque strate. La valeur de β ne dépend pas de la strate (pas d'interaction avec les facteurs non observés).

La régression logistique conditionnelle

On écrit la vraisemblance en conditionnant strate par strate, ce qui fait disparaître les α_k :

$$\ell(\beta) = \prod_{k=1}^{K} \ell_k(\beta),$$

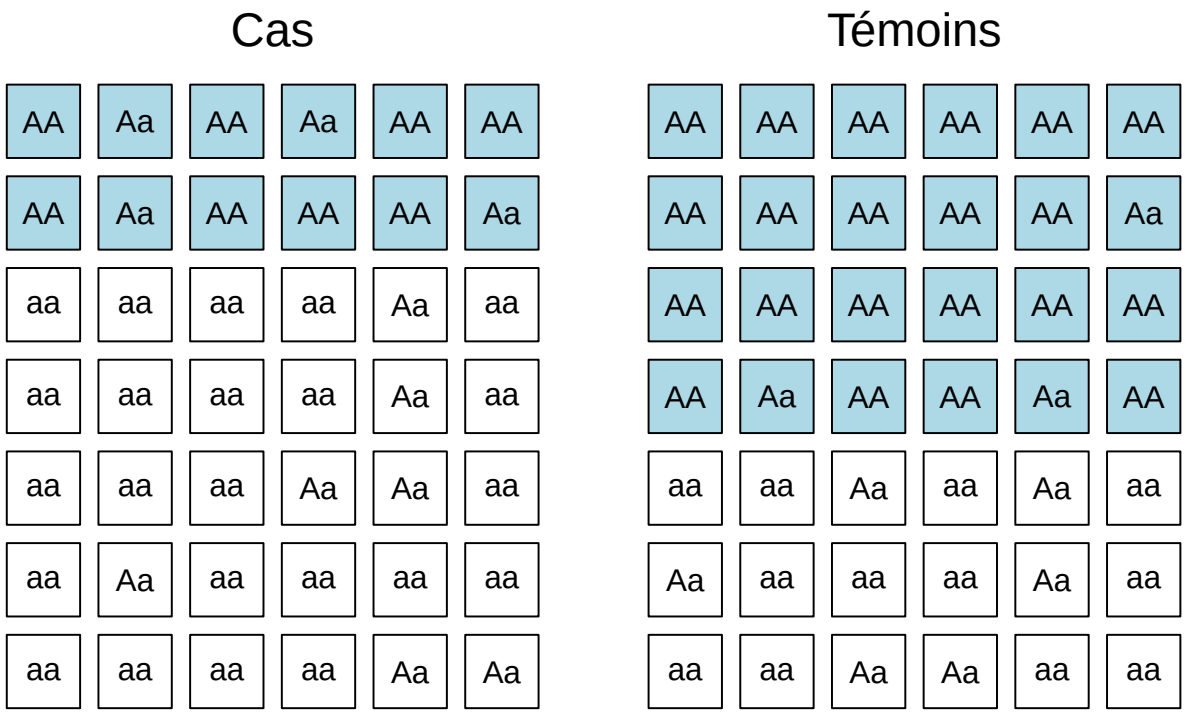
où $\ell_k(\beta)$ est la vraisemblance pour la strate k, écrite comme précédemment dans le cas d'une strate unique,

$$\ell_k(\beta) = \frac{e^{\sum_i \beta' x_{ki} y_{ki}}}{\sum_{v \in V} e^{\sum_i \beta' x_{ki} v_i}}.$$

La régression logistique conditionnelle

Exemple

On peut analyser ces données en faisant deux strates.



AA	Aa	aa
8	4	0
0	7	23

AA	Aa	aa
21	3	0
0	6	12

☞ test non significatif.

TDT et régression logistique conditionnelle

1. Le Transmission Disequilibrium Test (TDT)
2. Risques relatifs
3. La régression logistique conditionnelle (RLC)
4. Le TDT par la RLC

Pseudo-contrôles AFBAC

On revient au cadre du TDT : on considère un marqueur di-allélique A/a et des trios génotypés en ce marqueur.

Désignons par $\alpha\beta$ et $\gamma\delta$ les génotypes des parents d'un cas de génotype $\alpha\gamma$. Nous construisons trois génotypes qui correspondent à trois individus fictifs, appelés « pseudo-contrôles », en prenant un allèle de chacun des parents : $\alpha\delta$, $\beta\gamma$ et $\beta\delta$.

Par exemple, pour des parents AA et Aa avec un enfant AA, ces trois pseudo-contrôles ont les génotypes AA, Aa, Aa.

Pour le même appariement parental et un enfant Aa, les trois pseudo-contrôles ont les génotypes AA, AA, Aa.

Pseudo-contrôles AFBAC

Appariement	Génotype du cas	Génotypes des pseudo-contrôles		
AA × AA	AA	AA	AA	AA
AA × Aa	AA	AA	Aa	Aa
	Aa	AA	AA	Aa
AA × aa	Aa	Aa	Aa	Aa
Aa × Aa	AA	Aa	Aa	aa
	Aa	AA	Aa	aa
	aa	AA	Aa	Aa
Aa × aa	Aa	Aa	aa	aa
	aa	Aa	Aa	aa
aa × aa	aa	aa	aa	aa

Le TDT par la RLC

On va appliquer la régression logistique conditionnelle en créant une strate par famille, avec 4 individus : le cas (individu $i = 1$), et les trois pseudo-contrôles ($i = 2, 3, 4$).

Les génotypes de l'individu i sont recodés par $X_i = (X_{i1}, X_{i2})$.

génotype	X_{i1}	X_{i2}	$e^{\beta' X_i}$
AA	0	0	1
Aa	1	0	$e^{\beta_1} = \psi_1$
aa	0	1	$e^{\beta_2} = \psi_2$

La vraisemblance de la strate k, avec un cas et trois contrôles, est

$$\ell_k(\beta) = \frac{e^{\beta' x_1}}{e^{\beta' x_1} + e^{\beta' x_2} + e^{\beta' x_3} + e^{\beta' x_4}}.$$

Le TDT par la RLC

Reprenons l'exemple du trio AA × Aa avec un enfant AA, et trois pseudo-contrôles AA, Aa, Aa.

$$\ell_k(\beta) = \frac{e^{\beta' x_1}}{e^{\beta' x_1} + e^{\beta' x_2} + e^{\beta' x_3} + e^{\beta' x_4}} = \frac{1}{1+1+\psi_1+\psi_1} = \frac{1}{2} \times \frac{1}{1+\psi_1}.$$

Même appariement avec enfant Aa, trois pseudo-contrôles AA, AA, Aa :

$$\ell_k(\beta) = \frac{e^{\beta' x_1}}{e^{\beta' x_1} + e^{\beta' x_2} + e^{\beta' x_3} + e^{\beta' x_4}} = \frac{\psi_1}{\psi_1+1+1+\psi_1} = \frac{1}{2} \times \frac{\psi_1}{1+\psi_1}.$$

☞ On retrouve pour chaque strate une vraisemblance proportionnelle à celle de Schaid et Sommer.

Le TDT par la RLC

- Le conditionnement sur la strate formée par le cas et les trois pseudos contrôles AFBAC est équivalent au conditionnement sur les génotypes parentaux.
- Comme on l'a mentionné dans la partie 2, on retrouve exactement le TDT si on choisit le modèle additif.
- La souplesse du cadre de travail logistique facilite l'« ajustement » sur des covariables, la création de tests d'interaction, la sélection de variables, etc.
- Malheureusement les choses se compliquent sérieusement si on souhaite modéliser des haplotypiques (Cordell & Clayton, 2002).

Quelques références

- Spielman RC, McGinnis RE, Ewens WJ (1993). Transmission test for linkage disequilibrium : the insulin gene region and insulin-dependent diabetes mellitus. Am J Hum Genet 52.
- Schaid DJ, Sommer SS (1993). Genotype relative risks : methods for design and analysis of candidate-gene association studies. Am J Hum Genet 53.
- Cordel HJ, Clayton DG. (2002) A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms within a Gene Using Case/Control or Family Data : Application to HLA in Type 1 Diabetes. Am J Hum Genet 70.