CAPÍTOL 6

EL PROBLEMA DEL PUNT DE CANVI

Índex Capítol

CAPÍTOL 6

EL PROBLEMA DEL PUNT DE CANVI

En aquest capítol s'introdueix el problema del punt de canvi per a seqüències de dades normals, binàries i politòmiques. Les tècniques proposades seran utilitzades per resoldre el problema de l'homogeneïtat d'estil en un text literari, doncs permeten determinar si hi ha algun canvi en les distribucions de variables que caracteritzen l'estil, i localitzar el punt on canvien.

6.1 El problema del punt de canvi

Es diu que una seqüència de variables aleatòries independents $y_1, y_2, \ldots, y_n$ té un punt de canvi a $r^* \in \{1,2,\ldots n\}$ si y_i per $i \in \{1,2,\ldots r^*\}$ tenen per funció de distribució $F_a(y)$, mentre que y_i per $i \in \{r^*+1,\ldots,n\}$ tenen per funció de distribució $F_d(y)$, essent $F_a(y) \neq F_d(y)$. Tant el valor de r^* com els paràmetres que descriuen les funcions de distribució $F_a(y)$ i $F_d(y)$ en principi són desconeguts. D'aquí en endavant, i per a simplificar la notació, farem servir r en lloc de r^* per denotar el punt de canvi.

Quan la variable en estudi és la característica de qualitat d'un cert procés de producció, el canvi pot ser degut, per exemple, a un desajust sobtat del procés o a la introducció d'una nova partida de matèria prima. La detecció d'un punt de canvi a l'estudi de l'homogeneïtat d'estil literari pot permetre descobrir canvis d'autor o d'època en la que va ser escrit un text.

El problema de l'estimació del punt de canvi ha sigut estudiat de forma extensiva en la literatura. Per tal de determinar el valor de r en seqüències univariants s'han desenvolupat estimadors tant paramètrics com no paramètrics, tant en l'aproximació clàssica, via mètode màxim versemblant entre d'altres, com bayesiana. També s'han

considerat extensions al problema per seqüències multivariants, en els que, en la majoria dels casos, s'assumeix que les variables tenen distribucions contínues.

En el cas de variable contínua, siguin $f_a(y)$ i $f_d(y)$ les densitats de probabilitat abans i després de r, indexades pels paràmetres θ_a i θ_d respectivament. En el cas de variable discreta siguin les corresponents distribucions de probabilitat. Hinkley (1970) va enfocar el problema considerant l'estimació màxim versemblant del punt de canvi en una seqüència de variables contínues, amb èmfasi en el cas normal i canvi en la mitjana. La funció de versemblança del model proposat és:

$$l(r,\theta_a,\theta_d)=\prod_{i=1}^{r} f_a(y_i)\prod_{i=r+1}^{n} f_d(y_i).$$

L'estimador màxim versemblant de r, $\hat{r}$, és l'enter positiu r que maximitza el logaritme de la versemblança, $L(r,\theta_a,\theta_d)$, i que es defineix com a:

$$L(r,\theta_a,\theta_d)=\sum_{i=1}^{r}\log f_a(y_i)+\sum_{i=r+1}^{n}\log f_d(y_i),$$

si $r=1,2,...,n-1$, i per

$$L(r,\theta_a,\theta_d)=\sum_{i=1}^{n}\log f_a(y_i).$$

si $r = n$, o, dit d'una altra manera, si no hi ha canvi.

Hinkley (1970) va aplicar propietats dels random walks per derivar la distribució assimptòtica dels estimadors màxim versemblants.

6.1.1 Seqüència de Normals

Sen i Srivastava (1975 a,b,c) mostren com per y_i distribuïda *normalment* amb variança unitària, l'estadístic de raó de versemblança per contrastar la hipòtesi nul·la de que no hi ha canvi en front a l'alternativa de que sí que n'hi ha és equivalent a:

$$U_r=\frac{\frac{1}{r}\sum_{i=1}^{r} y_i-\frac{1}{n-r}\sum_{i=r+1}^{n} y_i}{\left(\frac{1}{r}+\frac{1}{n-r}\right)^{\frac{1}{2}}},$$

i estimen r com el valor r que fa màxim U_r, mentre que si la variança és desconeguda però constant, cal dividir l'estadístic per la variança mostral. També demostren que U_r és equivalent l'estadístic t usat per a testejar si hi ha diferències entre les mitjanes de les observacions anteriors i posteriors al punt de canvi. Hawkins (1977) i Worsley (1979) donen les distribucions exactes de l'estadístic quan la variança és coneguda i quan és desconeguda. A la secció 6.2 detallem la nostra aproximació a aquest problema.

6.1.2 Seqüència de binomials

Per a seqüències de *dades binàries* independents Pettitt (1980) suggereix l'estadístic:

$$T_r^{'} = n\left|\sum_{i=1}^{r} y_i - \frac{r}{n}\sum_{i=1}^{n} y_i\right|,$$

per testejar la hipòtesi de que no existeix canvi enfront a de que sí que n'hi ha, i estima r com el valor r que maximitza T_r'. Smith (1975) dóna estimadors bayesians per probabilitats a priori iguals per tots els potencials punts de canvi. Wolfe i Chen (1990) proposen estimadors que són fàcils de calcular i, en alguns casos, millors que els màxim versemblants. Donada una seqüència de n variables aleatòries Bernouilli mútuament independents, per a cada possible punt de canvi r, l'estimador màxim versemblant de $\Delta=\pi_d-\pi_a$ és:

$$T_r = \sum_{i=r+1}^{n}\left(\frac{y_i}{n-r}\right) - \sum_{i=1}^{r}\left(\frac{y_i}{r}\right),$$

que correspon a la diferència de les mitjanes mostrals de les dues subseqüències formades per les observacions $y_1, y_{2, \ldots,} y_r$ i $y_{r+1}, y_{r+2, \ldots,} y_n$, respectivament. Un estimador màxim versemblant del punt de canvi r és:

$$\hat{r}_{BMV} = min\{k : T_k = \max_{1\le r\le n-1} T_r\}\ .$$

Observem que en tractar dades discretes, pot ser que el màxim de T_r es pugui donar en més d'un valor de r. Wolfe i Chen (1990) defineixen l'estimador com l'enter més petit pel que es té el màxim T_r. Com que la variança de T_r, per $1\le$ r $\le n-1$, i donat un punt de canvi r, és funció de r, Wolfe i Chen (1990) consideren l'estadístic modificat:

$$U_r = T_r\sqrt{r(n-r)/n}\ ,$$

que és equivalent a l'estadístic t usat en la comparació de mitjanes per dades no aparellades quan la variança és 1, i igual a l'estadístic proposat per Sen i Srivastava (1975a) per seqüència de Normals. Estimen r com el valor r que maximitzi U_r:

$$\hat{r}_{W0} = min_r\{k : U_k = \max_{1\le r\le n-1} U_r\}.$$

Hinkley i Hinkley (1970) i Smith (1980) consideraren els estimadors màxim versemblants i bayesians, respectivament, per a una seqüència de variables aleatòries binomials independents, i Worsley (1983) dóna estadístics de raó de versemblança per seqüències binomials i Poisson. A la secció 6.3 detallem la nostra aproximació a aquest problema.

6.1.3 Seqüència de multivariants

Problemes de punt de canvi per *seqüències multivariants* es donen en àrees com l'estudi de la proporció de productes classificats com a correctes, lleugerament defectuosos i defectuosos, l'estudi del punt en el que una determinada política comença a tenir impacte en índexs econòmics, o l'estudi del punt en que una determinada medicina

comença a fer efecte sobre un pacient en experiments clínics o el punt en el que hi ha un canvi d'estil literari a partir de les proporcions de paraules de diferents llargades.

Srivastava i Worsley (1986) van estudiar els test de raó de versemblança per canvis en la mitjana de dades Normals multivariants i varen demostrar que el mètode podia ser utilitzat també per a obtenir tests aproximats per canvis en les probabilitats de les files en taules de contingència i Booth i Smith (1982) proposen estimadors bayesians per a seqüències multivariants. Wolfe i Chen (1990) proposen estimadors del punt de canvi per a seqüències multinomials mitjançant una aproximació univariant al problema. A la secció 6.4 detallem la nostra aproximació a l'estimació d'un punt de canvi en seqüències multinomials.

6.1.4 Més d'un punt de canvi

Vostrikova (1981) suggereix que es poden estimar un nombre indeterminat de punts de canvi a través d'un procediment de segmentació binària: s'estima un punt de canvi en la seqüència i, en cadascuna de les dues subseqüències formades per les observacions anteriors i posteriors a aquest, es busca per separat un nou punt de canvi. El procés es repeteix fins que en cap subseqüència es pot estimar un nou punt de canvi, per un determinat nivell de significació. Demostra que pel cas multivariant aquest procediment estima de forma consistent tots els punts de canvi en la seqüència. A la secció 6.5 detallem la nostra aproximació al problema de detecció de més d'un punt de canvi.

L'obtenció dels estimadors màxim versemblants és un problema d'optimització, que en el cas de la localització del punt de canvi és d'optimització entera. Proposem una alternativa a l'optimització, basada en l'ajust de models, molt més intuïtiva i de fàcil aplicació, i que s'adapta a l'estimació d'un i de més d'un punt de canvi per a seqüències de Normals, binomials i multinomials.

6.2 Punt de canvi per seqüència de Normals

Sigui $y_1, y_2, \ldots, y_n$ una seqüència de variables aleatòries independents *normalment* distribuïdes. En l'estudi de l'homogeneïtat del *Tirant* estudiarem el cas en el que canvia el valor esperat, mentre que la variança es manté constant, i l'aplicarem a la llargada mitjana de paraula i als índexs de diversitat.

La funció de versemblança per a una seqüència de normals en la que a r la distribució passa de $N(\mu_a,\sigma^2)$ a $N(\mu_d,\sigma^2)$, és:

$$l\left(r,\mu_a,\mu_d,\sigma^2\right)=\left(\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\right)^n\prod_{i=1}^{r}\exp\left[-\frac{1}{2}\left(\frac{y_i-\mu_a}{\sigma}\right)^2\right]\prod_{i=r+1}^{n}\exp\left[-\frac{1}{2}\left(\frac{y_i-\mu_d}{\sigma}\right)^2\right],$$

i el seu logaritme:

$$L\left(r,\mu_a,\mu_d,\sigma^2\right)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{r}(y_i-\mu_a)^2+\sum_{i=r+1}^{n}(y_i-\mu_d)^2\right).$$

L'estimador màxim versemblant de r és l'enter positiu r que maximitza $L(r,\mu_a,\mu_d,\sigma^2)$.

6.2.1 Comparació de mitjanes

L'objectiu de la comparació de dues mitjanes és contrastar una hipòtesi nul·la H_0: $\mu_a=\mu_d$ en front d'una hipòtesi alternativa H_1: $\mu_a \neq \mu_d$, on, a diferència del problema del punt de canvi, μ_a i μ_d són les mitjanes de les dues poblacions perfectament definides. El mètode clàssic de resolució del test es basa en les hipòtesis de Normalitat de les mitjanes mostrals, homogeneïtat de variances, aleatorietat en les mostres i de independència entre observacions, i consisteix en fer servir l'estadístic t:

$$t = \frac{\bar{y}_a - \bar{y}_d}{s\sqrt{\frac{1}{n_a}+\frac{1}{n_d}}},$$

on $\bar{y}_a$ i $\bar{y}_d$ són les dues mitjanes mostrals, n_a i n_d les grandàries de les dues mostres i s la millor estimació de la desviació tipus poblacional, en el cas que $\sigma_a^2=\sigma_d^2$:

$$s = \sqrt{\frac{(n_a-1)s_a^2+(n_d-1)s_d^2}{n_a+n_d-2}},$$

essent s_a^2 i s_d^2 les variances mostrals per a les dues mostres. Si la hipòtesi nul·la és certa i les hipòtesis distribucionals acceptables, l'estadístic t té distribució *t-student* amb n_a+n_d-2 graus de llibertat.

Una forma equivalent de resoldre el problema passa per l'ajust d'un model de regressió lineal simple, on les respostes són les y_i i la variable explicativa és una variable indicadora Ind_{1i} que codifica el grup al que pertany una observació, prenent valor $Ind_{1i}=0$ per a les observacions en la mostra a i valor $Ind_{1i}=1$ per a les observacions en la mostra d. El model lineal que s'ajustarà, per mínims quadrats, és:

$$y_i \mid Ind_{i1} \sim N(E(y_i \mid Ind_{i1}) = \beta_0 + \beta_1 Ind_{i1}, \sigma^2).$$

Observem que $E(y_i|Ind_{1i}=0)=\beta_0=\mu_a$, i que $E(y_i|Ind_{1i}=1)=\beta_0+\beta_1=\mu_d$. Es pot comprovar que si b_1 és el pendent estimat per mínims quadrats, $b_1 = \bar{y}_a - \bar{y}_d$, $s_R^2 = s^2$ i, per tant,

$$H_0: \beta_1=0=\mu_d-\mu_a$$

es pot contrastar fent:

$$t = \frac{b_1}{s_{b_1}} = \frac{\bar{y}_a - \bar{y}_d}{s\sqrt{\frac{1}{n_a}+\frac{1}{n_d}}},$$

on s_{b1} és la desviació tipus estimada per b_1.

Quan no es pot considerar que $\sigma_a^2=\sigma_d^2=\sigma^2$, es pot ajustar el model

$$y_i \mid Ind_{i1} \sim N(\beta_0 + \beta_1 Ind_{i1}, \sigma_i^2).$$

ponderant les observacions proporcionalment a $1/\sigma_i^2$.

En la secció 6.2.2 proposem estimar el punt de canvi trobant per quina r la diferència entre les mitjanes de les observacions $1,2,...,r$ i $r+1, ..., n$ és més gran.

6.2.2 Estimació del punt de canvi via regressió lineal

Proposem estimar el punt de canvi en una seqüència de *normals* mitjançant l'ajust de *n-1* models de regressió lineal simple del tipus:

$$y_i \mid Ind_{i1} \sim N\left(E\left(y_i \mid Ind_{i1}\right) = \beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}, \sigma_i^2\right), \qquad (6.1)$$

on $Ind_{1i}^{(r)}$ és una variable indicadora que pren valor *0* per *i=1,2,...,r* i valor *1* per *i=r+1,...,n*. L'ajust dels models es fa ponderant les observacions proporcionalment a la inversa de la seva variança.

Per estimar *r*, ajustem el model lineal (6.1) *n-1* vegades, per $1 \le r \le n-1$, i escollim el valor de *r* que millor ajusta les dades, d'acord amb un criteri de bondat de l'ajust especificat. Nosaltres triem maximitzar l'estadístic *F* de la taula ANOVA. Sigui F_r l'estadístic *F* obtingut per a *r*. Llavors *r* és estimat com:

$$\hat{r}_N = \max_{1 \le r \le n-1} F_r,$$

Observem que F_r és igual al quadrat de l'estadístic que fem servir per contrastar la hipòtesi que no hi ha diferència entre els dos grups en què *r* divideix la seqüència, quan les dues poblacions tenen la mateixa variança:

$$F_r = t_r^2 = \frac{b_1^2}{s_{b_1}^2}$$

És a dir, el que proposem és equivalent a fer les *n-1* comparacions de les mitjanes de les dues mostres formades per les observacions *1,2,...,r* i *r+1,...,n* respectivament, per $1 \le r \le n-1$, a través del *t-test* clàssic per dades no aparellades. En aquest cas, doncs, $\hat{r}_N$ es pot escriure com:

$$\hat{r}_N = \max_{1 \le r \le n-1} \left| t_r \right|.$$

Observem que $\hat{r}_N$ coincideix amb l'estimador màxim versemblant del punt de canvi i, per tant:

$$\hat{r}_N = \hat{r}_{MV}.$$

Per al model (6.1) de regressió lineal simple amb errors normals i variança constant, la funció de versemblança per a totes les *n* observacions és:

$$l\left(r, \beta_0^{(r)}, \beta_1^{(r)}, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0^{(r)} - \beta_1^{(r)} Ind_{1i}^{(r)}\right)^2 \right),$$

i el logaritme de la versemblança és:

$$L\left(r, \beta_0^{(r)}, \beta_1^{(r)}, \sigma^2\right) = \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0^{(r)} - \beta_1^{(r)} Ind_{1i}^{(r)}\right)^2.$$

Per a una *r* donada, el màxim de la versemblança per $\beta^{(r)} = \left(\beta_0^{(r)}, \beta_1^{r)}\right)$, $L\left(\beta_0^{(r)}, \beta_1^{r)}, \sigma^2\right)$, s'obté minimitzant l'expressió:

$$\min_\beta L\left(\beta_0^{(r)}, \beta_1^{r)}, \sigma^2\right) = \min_\beta \left(\log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0^{(r)} - \beta_1^{(r)} Ind_{1i}^{(r)}\right)^2 \right),$$

i per tant, per a una r donada, l'estimador màxim versemblant per $\beta^{(r)} = \left(\beta_0^{(r)}, \beta_1^{r)}\right)$ minimitza la suma de quadrats residual. Com que:

$$F_r = \frac{s_T^2}{s_R^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - SQ_R}{\frac{SQ_R}{n-2}},$$

obtenir el model que fa màxim l'estadístic F_r és equivalent a obtenir el model que fa mínima la suma de quadrats residual i que fa màxima la versemblança. Per tant, també estimem el punt de canvi, r, com:

$$\hat{r}_N = \max_{1 \le r \le n-1} L\left(r, \hat{\beta}^{(r)}\right),$$

on $L\left(r, \hat{\beta}^{(r)}\right)$ és el logaritme de la versemblança avaluat a r, i $\hat{\beta}^{(r)}$ l'estimador mínim quadràtic i màxim versemblant, dels paràmetres del model.

Com a eina per a estimar el punt de canvi utilitzem el gràfic amb l'evolució la mesura de qualitat de l'ajust ($|t_r|$, F_r o $L\left(r, \hat{\beta}^{(r)}\right)$) en funció de r. Permet visualitzar el màxim que correspon a l'estimador $\hat{r}_N$ del punt de canvi. La forma del gràfic permet veure si el màxim és prou clar i definit, en el cas que existeixi un màxim identificable sense ambigüitats, i si existeixen altres possibles màxims locals que aportin claus a la interpretació de la frontera, com per exemple la possibilitat de que n'hi hagi més d'una. Aquests gràfics estan molt relacionats amb els *profile log-likelihood*.

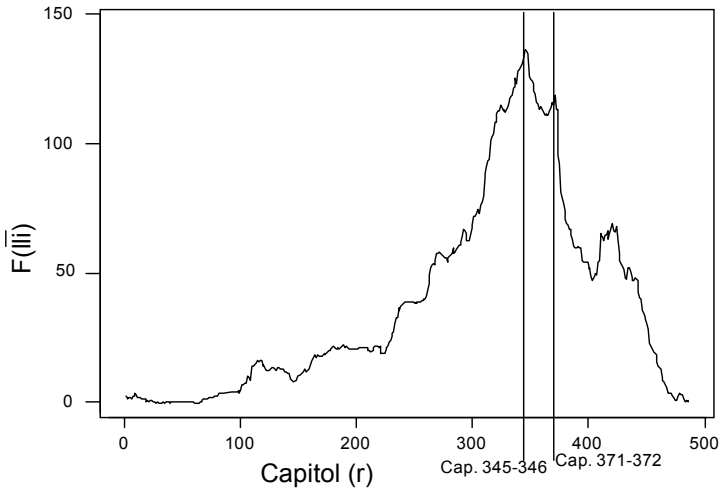

Figura 6.1: Gràfic de l'evolució de F_r en funció de r per a l'estimació del punt de canvi en la seqüència de llargades mitjanes de paraula per capítol. El màxim de F_r es té per *r=345*.

La figura 6.1 mostra el gràfic l'evolució de F_r en funció de r, per la seqüència de llargades mitjanes de paraula que s'estudiarà en el capítol 8, quan r és el nombre de capítols des del començament del llibre fins a la frontera en estudi. L'anàlisi del gràfic indica que:

- El màxim global de l'estadístic F_r es troba en el capítol *345*, que indica una possible frontera d'estil en aquest punt.
- El màxim és clar i ben definit.
- Hi ha un màxim local per *r=371*.

6.3 Punt de canvi per seqüència de binomials

Suposem que la variable aleatòria pot prendre només un de dos possibles valors com, per exemple, presència o absència d'un atribut, i sigui π_i la probabilitat de que una unitat presa a l'atzar en la mostra i-èssima tingui aquest atribut. Si extraiem mostres aleatòries de N_i observacions, el nombre d'unitats en les mostres que tenen aquest atribut, y_i, es distribueix segons una *Binomial*(N_i, π_i).

En l'estudi de l'homogeneïtat del *Tirant* estimarem el punt de canvi en cadascuna de les $l=10$ seqüències de binomials formades per les proporcions d'ocurrències de paraules de j lletres, per $j=1,2,...,10+$ en el capítol 8, en totes les $l=8$ seqüències de proporcions de frases en cadascuna de les categories en que s'han classificat segons la seva llargada en el capítol 9, i en cadascuna de les $l=25$ seqüències de proporcions d'ús de les paraules més abundants en el capítol 11.

La seqüència de binomials té un punt de canvi a r si $\pi_i=\pi_a$ per $i=1,2,...,r$ i $\pi_i=\pi_d$ per $i=r+1,...,n$. La funció de versemblança en aquest cas és:

$$l(r,\pi_a,\pi_d)=\prod_{i=1}^{r}\binom{N_i}{y_i}\pi_a^{y_i}\left(1-\pi_a^{y_i}\right)^{N_i-y_i}\prod_{i=r+1}^{n}\binom{N_i}{y_i}\pi_d^{y_i}\left(1-\pi_d^{y_i}\right)^{N_i-y_i}.$$

L'estimador màxim versemblant de r és l'enter positiu r que maximitza $l(r,\pi_a,\pi_d)$. El logaritme de la versemblança es defineix com:

$$L(r,\pi_a,\pi_d)=\sum_{i=1}^{r}\left(y_i\log\pi_a+(N_i-y_i)\log(1-\pi_a)+\log\binom{N_i}{y_i}\right)+$$
$$+\sum_{i=r+1}^{n}\left(y_i\log\pi_d+(N_i-y_i)\log(1-\pi_d)+\log\binom{N_i}{y_i}\right).$$

Quan la grandària de la mostra és constant, $N_i=N$ per a totes les mostres.

6.3.1 Comparació de dues proporcions

Comparar dues proporcions consisteix en contrastar la hipòtesi $H_0: \pi_a = \pi_d$ en front d'una hipòtesi alternativa $H_1: \pi_a \neq \pi_d$, on π_a i π_d són les probabilitats de que prenent una unitat a l'atzar de les poblacions a i d perfectament definides, tingui l'atribut. Si es prenen mostres de grandària N_i i s'hi compten y_i unitats amb un determinat atribut, la proporció d'unitats que posseeixen aquest atribut,

$$\hat{\pi}_i=\frac{y_i}{N_i}$$

és l'estimador màxim versemblant de la probabilitat π_i.

Sigui y_i una variable *binomial*(N_i, π_i). La comparació de proporcions es pot fer de forma anàloga a la comparació de mitjanes, ajustant models lineals generalitzats:

$$y_i|Ind_{1i} \sim Binomial\ (N_i, \pi_i=g^{-1}(\beta_0+\beta_1 Ind_{1i}))\ ,$$

on $g(\pi_i)$ és la funció *link*. Les funcions *link* habituals per a l'anàlisi de dades binomials que resolen aquest problema són:

- logit: $g(\pi_i) = \log\dfrac{\pi_i}{1-\pi_i}$,
- probit: $g(\pi_i) = \Phi^{-1}(\pi_i)$ on Φ^{-1} és la inversa de la funció de distribució de la Normal,
- complementary log-log: $g(\pi_i) = \log[-\log(1-\pi_i)]$ que és la inversa de la funció de distribució del valor extrem.

El model a ajustar per a la comparació de dues proporcions usant el *link logit* es:

$$y_i \mid Ind_{1i} \sim Binomial\left(N_i, \pi_i = \frac{\exp(\beta_0 + \beta_1 Ind_{1i})}{1+\exp(\beta_0 + \beta_1 Ind_{1i})}\right).$$

Triar com a criteri el maximitzar el logaritme de la versemblança és equivalent a minimitzar la *Deviança*, *D*, definida com:

$$D = -2\log\left[\frac{Màxima\ \ Versemblança\ \ del\ \ Model}{Màxima\ \ Versemblança\ \ del\ \ Model\ \ Saturat}\right].$$

El model suposa que el valor estimat pel model saturat coincideix amb l'observat. D'aquesta manera, la *Deviança* per dades binomials queda:

$$D = -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{N_i\hat{\pi}_i}{y_i}\right) + (N_i - y_i)\log\left(\frac{N_i - N_i\hat{\pi}_i}{N_i - y_i}\right)\right].$$

La *Deviança* juga en models lineals generalitzats el mateix paper que juga la suma de quadrats residuals en models lineals amb errors Normals, i com a tal és molt útil per a seleccionar el millor model entre els que tenen el mateix nombre de paràmetres, com és el nostre cas.

Un altre criteri d'ajust, consisteix en minimitzar l'estadístic X^2 de Pearson, que per la binomial és:

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - N_i\hat{\pi}_i)^2}{N_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

Aquest criteri és anàleg a minimitzar la suma de quadrats ponderada, amb la diferència que ara els pesos depenen dels paràmetres del model mentre que per la regressió per dades *normals* no, i és equivalent a minimitzar la distància χ^2 definida a (5.3).

De forma anàloga al que passa per la comparació de dues mitjanes, π_a-π_d=0 és equivalent a que β_1=0. El test de la raó de versemblança per contrastar aquesta hipòtesi fa servir l'estadístic:

$$F = \frac{\dfrac{D_0 - D_1}{1}}{\dfrac{D_1}{N_a + N_d - 2}} \sim F(1, N_a + N_d - 2)\ .$$

on D_1 i D_0 són les *Deviances* per al model complet, amb $\beta_1 \neq 0$, i el model simplificat, amb β_1=0, respectivament, i on N_a i N_d són les grandàries de les dues mostres. Si s'obtenen valors de *F* grans, relatius a la distribució $F(1,N_a+N_d-2)$, hi ha evidències a

favor de rebutjar H_0 i, per tant, que hi ha diferències entre les proporcions. Assimptòticament X^2 és equivalent a la Deviança i, per tant, pel que fa a la inferència valen les mateixes consideracions.

Per mostres suficientment grans la distribució binomial es pot aproximar per una Normal. En aquest cas el test aproximat de comparació dues proporcions, π_a i π_d, és el mateix que es realitza per comparar dues mitjanes:

$$t = \frac{\hat{\pi}_a - \hat{\pi}_d}{\sqrt{\frac{\hat{\pi}_a(1-\hat{\pi}_a)}{N_a} + \frac{\hat{\pi}_d(1-\hat{\pi}_d)}{N_d}}} .$$

S'obtenen idèntics resultats ajustant models lineals:

$$\frac{y_i}{N_i} \mid Ind_{1i} \sim N\left(\pi_i = \beta_0 + \beta_1 Ind_{1i}, \frac{\pi_i(1-\pi_i)}{N_i}\right) ,$$

on y_i és el nombre d'ocurrències en la mostra i-èssima i Ind_{1i} és la variable indicadora que pren valor *0* per $i \in a$ i valor *1* per $i \in d$, assignant a les observacions pesos proporcionals a l'invers de la seva variança.

6.3.2 Estimació del punt de canvi via regressió logística

De forma anàloga al cas Normal, proposem estimar el punt de canvi en una seqüència binomial, $y_1,y_2,...y_n$, mitjançant l'ajust de *n-1* models logístics del tipus:

$$y_i \sim Binomial\left(N_i, \pi_i = \frac{\exp\left(\beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}\right)}{1+\exp\left(\beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}\right)}\right), \qquad (6.2)$$

on N_i és la grandària de la mostra i $Ind_{1i}^{(r)}$ és la variable indicadora que pren valor *0* per *i=1,2,...,r* i valor *1* per *i=r+1,...,n*. Comparem els *n-1* models ajustats sota l'estimació màxim versemblant dels seus paràmetres, i ens quedem aquell que maximitza $L\left(r, \hat{\beta}^{(r)}\right)$. És a dir, estimem *r* com a:

$$\hat{r}_L = \min\left\{k : L\left(k, \hat{\beta}^{(k)}\right) = \max_{1\le r\le n-1} L\left(r, \hat{\beta}^{(r)}\right)\right\} .$$

El que es proposa és equivalen a trobar el punt per el que la diferència entre les proporcions per a les dues subseqüències $y_1,y_2,...,y_r$ *i* $y_{r+1},...,y_n$ és més significativa. Quan utilitzem una altra de les funcions links plantejades en l'apartat 6.3.1, els resultats que s'obtenen, pel que fa a log-versemblança i probabilitats estimades pel model plantejat, són coincidents amb les obtingudes pel *link logit*. Observem, a més a més, que com ja passava pel cas normal, $\hat{r}_L$ coincideix amb l'estimador màxim versemblant del punt de canvi.

De manera anàloga al cas normal en problemes d'estimació del punt de canvi presentem visualitzem amb el gràfic de $L\left(r, \hat{\beta}^{(r)}\right)$ en funció de *r*, i el màxim d'aquesta corba correspon a l'estimador $\hat{r}_L$ del punt de canvi. El gràfic de la figura 6.2 mostra l'evolució de $L\left(r, \hat{\beta}^{(r)}\right)$ en funció de *r*, per la seqüència d'ocurrències de paraules de 10 i més

lletres, quan *r* és el nombre de capítols des del començament del llibre fins a la frontera en estudi. L'anàlisi del gràfic indica que:

- Hi ha un màxim de l'estadístic $L\left(r,\hat{\beta}^{(r)}\right)$ en el capítol 345, que indica una possible frontera d'estil en aquest punt.
- El màxim és únic i està definit sense ambigüitats.

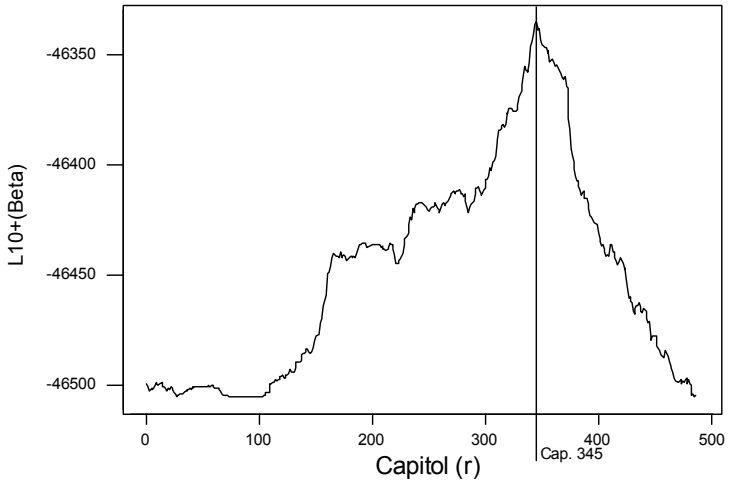

Figura 6.2: Evolució de $L\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r* per la seqüència d'ocurrències de paraules de 10 i més lletres en els capítols del *Tirant*. L'estimació del punt de canvi és $\hat{r}_L^*$=345.

Quan l'aproximació Normal és prou bona, també podriem estimar el punt de canvi en una seqüència binomial, $y_1,y_2,...y_n$, mitjançant l'ajust del model lineal:

$$\frac{y_i}{N_i} \sim N\left(\pi_i = \beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}, \frac{\pi_i(1-\pi_i)}{N_i}\right), \qquad (6.3)$$

per *r=1,2,...,n-1*. L'ajust dels models s'hauria de fer ponderant les observacions proporcionalment a la inversa de la seva variança. En la pràctica, les probabilitats π_i són desconegudes, però com que les diferències entre π_a i π_d són petites respecte de les diferències entre les N_i es pot suposar que els pesos són proporcionals a N_i.

Per estimar *r*, ajustem el model (6.3) *n-1* vegades i prenem el valor de *r* que millor ajusta les dades. Triant l'estadístic *F* de la taula ANOVA, *r* és estimat com:

$$\hat{r}_{NB} = \max_{1\le r\le n-1} F_r .$$

6.4 Punt de canvi en seqüència de multinomials

Donada una seqüència $\boldsymbol{y_i}=(y_{1i},y_{2i},....,y_{li}) \sim Mult(N_i,\boldsymbol{\pi_i}=(\pi_{1i},\pi_{2i},...,\pi_{li}))$, pot haver-hi un punt, *r*, en el que canvin les probabilitats que una unitat presa a l'atzar sigui classificada en algunes o en totes les categories, passant de $\boldsymbol{y_i} \sim Mult(N_i;\boldsymbol{\pi_a}=(\pi_{1a},\pi_{2a},...,\pi_{la}))$ a $\boldsymbol{y_i} \sim Mult(N_i;\boldsymbol{\pi_d}=(\pi_{1d},\pi_{2d},...,\pi_{ld}))$.

En la tesi estimarem el punt de canvi per les seqüències de multinomials formades per les ocurrències de paraules de *k* lletres, per *k=1,2,...,10+*, pels nombres de frases en

cadascuna de les vuit categories en que s'han classificat segons la seva llargada, i pel nombre d'ocurrències de les 25 paraules més abundants.

L'estimador màxim versemblant del punt de canvi s'obté maximitzant la funció de versemblança:

$$l(r,\pi_a,\pi_d)=\prod_{i=1}^{r}\frac{N_i!}{y_{1i}!y2_{2i}!...y_{li}!}\pi_{1a}^{y_{1i}}\pi_{2a}^{y_{2i}}...\pi_{la}^{y_{li}}\prod_{i=r+1}^{n}\frac{N_i!}{y_{1i}!y2_{2i}!...y_{li}!}\pi_{1d}^{y_{1i}}\pi_{2d}^{y_{2i}}...\pi_{ld}^{y_{li}},$$

o, de forma equivalent, de maximitzar el logaritme de la versemblança:

$$L(r,\pi_a,\pi_d)=\sum_{i=1}^{n}\log\left(\frac{N_i}{y_{1i}!y_{2i}!...y_{li}!}\right)+\sum_{i=1}^{r}\left(y_{1i}\log\pi_{1a}+...+y_{li}\log\pi_{la}\right)+$$
$$+\sum_{i=r+1}^{n}\left(y_{1i}\log\pi_{1d}+...+y_{li}\log\pi_{ld}\right).$$

Els estimadors màxim versemblants de π_a i π_d són:

$$\hat{\pi}_{ja}=\frac{\sum_{i=1}^{r}y_{ji}}{\sum_{i=1}^{r}N_i}\quad\text{i}\quad\hat{\pi}_{jd}=\frac{\sum_{i=r+1}^{n}y_{ji}}{\sum_{i=r+1}^{n}N_i}.$$

6.4.1 Estimació del punt de canvi per regressió politòmica

Amb poques variacions, la metodologia de detecció del punt de canvi proposada a la secció 6.3.2 per a una seqüencia de binomials pot ser adaptada per a respostes multinomials, amb més de dues categories. Per a la metodologia d'ajust de models per dades politòmiques, veure McCullagh i Nelder (1983) o Hosmer i Lemeshow (1989). L'estimació del punt de canvi que s'obté és la mateixa que pel mètode de la màxima versemblança.

Per respostes multinomials, hi ha una extensió natural del model (6.2). Fent servir la funció *link logit* multinomial tenim:

$$g(\pi_i)=(g_2(\pi_i),....,g_l(\pi_i)),$$

on:

$$g_j(\pi_i)=\log\left(\frac{\pi_{ji}}{\pi_{1i}}\right)=\beta_{0j}^{(r)}+\beta_{1j}^{(r)}Ind_{1i}^{(r)},\qquad(6.4)$$

per *j=2,...,l*. Prenent com a referència la categoria *1* i per a una *r* donada, podem modelar y_i ajustant el model (6.4) pel mètode de la màxima versemblança.

Recordem que $Ind_{1i}^{(r)}$ és una variable indicadora que pren valor *0* per les mostres *1,2,...,r* i valor *1* per les mostres *r+1,...,n*. Observem que aquest model té *2×(l-1)* paràmetres, $\beta^{(r)}=\left(\beta_{20}^{(r)},\beta_{30}^{(r)},...,\beta_{l0}^{(r)},\beta_{21}^{(r)},...,\beta_{l1}^{(r)}\right)$, i que el *link logit* multinomial és una generalització de la funció *link logit* emprada per dades binomials.

El model logístic per respostes binàries és un cas particular de (6.4), en el que *l=2* i on, per tant, $\pi_{2i}=1-\pi_{1i}$.

Comparem els *n-1* models ajustats sota l'estimació màxim versemblant dels seus paràmetres, i ens quedem aquell que maximitza $L\left(r,\hat{\beta}^{(r)}\right)$. És a dir, estimem *r* com a:

$$\hat{r}_M = \min\left\{k : L\left(k,\hat{\beta}^{(k)}\right) = \max_{1\le r\le n-1} L\left(r,\hat{\beta}^{(r)}\right)\right\}$$

La funció de versemblança condicional del model per a una mostra de *n* observacions independents és:

$$l(r,\boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_1 \left(Ind_{1i}^{(r)}(i)\right)^{y_{1i}} \pi_2 \left(Ind_{1i}^{(r)}(i)\right)^{y_{2i}} \ldots \pi_l \left(Ind_{1i}^{(r)}(i)\right)^{y_{li}} .$$

Prenent el logaritme de la versemblança, aprofitant el fet que:

$$\sum_{j=1}^{l} y_{ji} = N_i \text{ , i que } \sum_{j=1}^{l} \pi_{ji} = 1,$$

per a totes les *i=1,2,...,n*, i prenent la categoria *1* com a referència, la log-versemblança a maximitzar queda:

$$L\left(r,\beta^{(r)}\right) = \sum_{i=1}^{n}\left(\sum_{j=2}^{l} y_{ji}\left(\beta_{0j}^{(r)} + \beta_{1j}^{(r)} Ind_{1i}^{(r)}\right) - \ln\left(1 + \sum_{j=2}^{l} e^{\left(\beta_{0j}^{(r)} + \beta_{1j}^{(r)} Ind_{1i}^{(r)}\right)}\right)\right),$$

que és la mateixa que havíem obtingut per l'estimador màxim versemblant. Per tant, el valor estimat per a *r* amb aquest mètode coincideix amb l'estimador màxim versemblant.

Hem pres la categoria *1* com a referència, i modelem el logaritme de la relació entre les probabilitats de les *j-1* categories i la de referència per a *r=1,2,...,n-1*. S'obtenen resultats idèntics, pel que fa al logaritme de la versemblança, prenent qualsevol altre categoria com a referència.

Un altre criteri d'ajust dels models lineals generalitzats, tal i com s'ha descrit en la secció 6.3.1, consisteix en minimitzar l'estadístic X^2 de Pearson, que per la multinomial és equivalent a la distància χ^2:

$$X^2 = \sum_{i=1}^{n}\sum_{j=1}^{l} \frac{\left(y_i - N_i\hat{\pi}_{ji}\right)^2}{N_i\hat{\pi}_{ji}}.$$

Com havíem fet per dades binàries, presentarem els resultats a través del gràfic de $L\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r*. El gràfic permet visualitzar el màxim, que correspon a l'estimador $\hat{r}_M$ del punt de canvi, pel que tornen a valer totes les consideracions fetes anteriorment.

6.4.2 Aproximació univariant al cas multinomial

Quan no es disposa de mitjans computacionals que permeten ajustar models politòmics, es pot tractar el model multinomial mitjançant aproximacions univariades, considerant *l-1* seqüències de binomials.

Partint del fet que per a una categoria *j* donada, y_{ji} té una distribució *binomial*(N_i, π_{ja}) per $i=1,2,... r$, i *binomial*(N_i, π_{jd}) per $i=r+1,...n$, i suposant que hi ha com a màxim un punt de canvi a *r,* Wolfe i Chen (1990) considerant els estadístics:

$$W_{j(r)} = \sqrt{\frac{1}{N}}\sqrt{\frac{r(n-r)}{n}}\left(\sum_{i=r+1}^{n}\frac{y_{ji}}{(n-r)} - \sum_{i=1}^{r}\frac{y_{ji}}{r}\right),$$

per a una seqüència de multinomials amb *N* constant. Proposen estimar *r* a través de:

$$\hat{r}_{W1} = \min\left\{k : W_{(k)} = \max_{1\le r\le n-1}\max_{1\le j\le l}\left|W_{j(r)}\right|\right\}, \tag{6.5}$$

$$\hat{r}_{W2} = \min\left\{k : \sum_{j=1}^{l}\left|W_{j(k)}\right| = \max_{1\le r\le n-1}\sum_{j=1}^{l}\left|W_{j(r)}\right|\right\}, \tag{6.6}$$

$$\hat{r}_{W3} = \min\left\{k : \sum_{j=1}^{l}W_{j(k)}^{2} = \max_{1\le r\le n-1}\sum_{j=1}^{l}W_{j(r)}^{2}\right\}. \tag{6.7}$$

Arriben a la conclusió que $\hat{r}_{W3}$ és la millor estimació quan no hi ha gaire diferència entre les components de π_a i π_d o be quan cap de les components de π_a i π_d és molt petita. L'estimador màxim versemblant, pel contrari, és el millor quan cap de les dues condicions anteriors s'acompleix.

Proposem, com a aproximació a l'estimació del punt de canvi en una seqüència de multinomials, adaptar les propostes de Wolfe i Chen ajustant models logístics (6.2) per a cadascuna de les *l* categories, i substituir a (6.5), (6.6) i (6.7) $W_{j(r)}$ per $L_j\left(r, \hat{\beta}^{(r)}\right)$, el màxim del logaritme de la versemblança per al model ajustat per a la categoria j-èssima i per a una *r* donada. Els tres estimadors de *r* que proposem són:

$$\hat{r}_{ML1} = \min\left\{k : L_k\left(k, \hat{\beta}^{(k)}\right) = \max_{1\le r\le n-1}\max_{1\le j\le l}L_j\left(r, \hat{\beta}^{(r)}\right)\right\}, \tag{6.8}$$

$$\hat{r}_{ML2} = \min\left\{k : \sum_{j=1}^{l}L_j\left(k, \hat{\beta}^{(k)}\right) = \max_{1\le r\le n-1}\sum_{j=1}^{l}L_j\left(r, \hat{\beta}^{(r)}\right)\right\}, \tag{6.9}$$

$$\hat{r}_{ML3} = \min\left\{k : \sum_{j=1}^{l}L_j^2\left(k, \hat{\beta}^{(k)}\right) = \min_{1\le r\le n-1}\sum_{j=1}^{l}L_j^2\left(r, \hat{\beta}^{(r)}\right)\right\}, \tag{6.10}$$

on $\hat{\beta}^{(r)} = \left(\hat{\beta}_{20}^{(r)}, \hat{\beta}_{30}^{(r)}, ..., \hat{\beta}_{l0}^{(r)}, \hat{\beta}_{21}^{(r)}, ..., \hat{\beta}_{l1}^{(r)}\right)$ són els estimadors màxim versemblants dels paràmetres del model.

Les expressions (6.7) i (6.10) no coincideixen perquè, mentre el valor de $W_{j(r)}$ és sempre positiu, el màxim del logaritme de la versemblança, $L_j\left(r, \hat{\beta}^{(r)}\right)$, és negatiu. L'anàleg a obtenir el màxim de $L_j\left(r, \hat{\beta}^{(r)}\right)$ és trobar el mínim de $L_j^2\left(r, \hat{\beta}^{(r)}\right)$.

Els estadístics que proposem en aquest cas s'assemblen molt als de Wolfe i Chen. Tenen, però, l'avantatge de ser aplicables també quan les grandàries de mostra, N_i, no són constants.

Com havíem en tots els casos anteriors, podem visualitzar els resultats a través del gràfic de l'estadístic a maximitzar en funció de r. Per exemple, per visualitzar $\hat{r}_{ML2}$ busquem el màxim en el gràfic de $\sum_{j=1}^{l} L_j\left(r, \hat{\beta}^{(r)}\right)$ en funció de r. Per aquests gràfics valen totes les consideracions fetes anteriorment.

6.5 Estimació de més d'un punt de canvi

Fins a aquest punt hem suposat que hi ha un sol punt de canvi en una seqüència, tant per dades normals, com binàries o politòmiques. El mètode proposat té una extensió natural quan la seqüència analitzada té més d'un punt de canvi.

Sigui la seqüència $y_1, y_2, ..., y_n$, on les y_i són independents. Hi haurà $p-1$ punts de canvi localitzats a $r_1^*, r_2^*, ... , r_{P-1}^* \in \{1,2,...,n\}$ si per $i \in \{1,2,...,r_1^*\}$ y_i tenen per funció de distribució $F_1(y)$, per $i \in \{r_1^*+1,...,r_2^*\}$ tenen una funció de distribució $F_2(y)$, i així successivament fins a y_i per $i \in \{r_{p-1}^*+1,..., n\}$ tenen una funció de distribució $F_p(y)$, amb $F_1(y) \neq F_2(y)$, $F_2(y) \neq F_3(y)$, ..., $F_{p-1}(y) \neq F_p(y)$. Els valors de r_k^*, per $k=1,...,p-1$ i els paràmetres que descriuen les funcions de distribució $F_k(y)$, θ_k, són desconeguts.

El logaritme de la versemblança, pel cas general és:

$$L\left(r_1,...,r_{p-1},\theta_1,...,\theta_p\right) = \sum_{i=1}^{r_1} \log f_1(y_i) + \sum_{i=r_1+1}^{r_2} \log f_2(y_i) + ... + \sum_{i=r_{p-1}+1}^{n} \log f_p(y_i),$$

on, $f_k(y)$ són les p densitats de probabilitat o les distribucions de probabilitat en el cas de variable contínua o discreta, respectivament.

En l'apartat 6.5.1 es repassa la comparació de més de dues mitjanes o proporcions, i en el 6.5.2 s'exemplificarà una manera d'estimar els $p-1$ punts de canvi sobre el cas d'una seqüència de dades binomials.

6.5.1 Comparació de més de dues mitjanes o proporcions

De forma anàloga a com s'havia fet per a la comparació de 2 mitjanes, una manera de comparar p mitjanes, quan les p poblacions estan perfectament definides, es basa en l'ajust de models de regressió lineal, on les respostes són les y_i, i les $p-1$ variables explicatives són variables indicadores Ind_{ki} que codifiquen el grup al que pertany una observació, prenent valor $Ind_{ki}=1$ per a les observacions en la mostra k-èssima i 0 altrament, per $k=1,...,p-1$. S'ajusta el model lineal:

$$y_i | Ind_{ki} \sim N\left(\beta_0+\beta_1 Ind_{1i}+....+\beta_{p-1} Ind_{p-1i}, \sigma_i^2\right), \qquad \text{independents.}$$

Quan $Var(y_i|Ind_{ki})=\sigma_i^2$ no és constant, però es manté la independència, s'ajusta aquest model per mínims quadrats ponderats, amb pesos $w_i=C/\sigma_i^2$. La interpretació dels

paràmetres del model és anàloga a la vista per la comparació de dues mitjanes: $\beta_0=\mu_p$ i $\beta_k=\mu_k-\mu_p$ per $k=1,...,p-1$. Per tant, comprovar la hipòtesi de que totes les mitjanes poblacionals són iguals és el mateix que provar la hipòtesi que el vertader model és:

$$y_i|Ind_{ki} \sim N(\beta_0, \sigma_i^2) .$$

L'estadístic F de la taula *ANOVA* permet avaluar si es tenen evidències per rebutjar aquesta hipòtesi.

De forma anàloga, es pot realitzar la comparació de p proporcions. Es vol contrastar la hipòtesi de que les proporcions π_k són iguals, per $k=1,...,p$, en front a l'alternativa de que al menys una no ho és. S'ajusta el model amb $p-1$ variables indicadores:

$$y_i \mid Ind_{ki} \sim Binomial\left(N_i, \pi_i = g^{-1}(\beta_0 + \beta_1 Ind_{1i} + ... + \beta_{p-1} Ind_{p-1i})\right),$$

que amb el *link logit* queda com:

$$y_i \mid Ind_{ki} \sim Binomial\left(N_i, \pi_i = \frac{\exp\left(\beta_0 + \beta_1 Ind_{1i} + ... + \beta_{p-1} Ind_{p-1i}\right)}{1+\exp\left(\beta_0 + \beta_1 Ind_{1i} + ... + \beta_{p-1} Ind_{p-1i}\right)} \right).$$

De forma anàloga al que passa per la comparació de dues proporcions:

$$H_0: \pi_1 = \pi_2 = .. . = \pi_p$$

és equivalent a:

$$H_0: \beta_1 = \beta_2 = ... = \beta_{p-1} = 0.$$

6.5.2 Estimació de més d'un punt de canvi per seqüència de binomials

Sigui la seqüència $y_1, y_2, ..., y_n$, on les y_i són independents distribuïdes segons la binomial i on hi ha $p-1$ punts de canvi a $\boldsymbol{r}=(r_1,r_2,...,r_{P-1})$. El logaritme de la versemblança és:

$$\begin{aligned} L\left(r_1,...,r_{p-1},\pi_1,...,\pi_p\right) = & \sum_{i=1}^{r_1}\left(y_i \log\pi_1 + \left(N_i - y_i\right)\log\left(1-\pi_1\right) + \log\binom{N_i}{y_i}\right) + \\ & + \sum_{i=r_1+1}^{r_2}\left(y_i \log\pi_2 + \left(N_i - y_i\right)\log\left(1-\pi_2\right) + \log\binom{N_i}{y_i}\right) + ... + \\ & + \sum_{i=r_{p-1}+1}^{n}\left(y_i \log\pi_p + \left(N_i - y_i\right)\log\left(1-\pi_p\right) + \log\binom{N_i}{y_i}\right). \end{aligned}$$

L'estimador màxim versemblant de $\boldsymbol{r}$ s'obté maximitzant el logaritme de la versemblança.

Si el nombre de punts de canvi, $p-1$, és conegut, un mètode d'estimació de les localitzacions consisteix en ajustar models lineals, prenent y_i com a resposta, i com a explicatives les variables indicadores $Ind_{ki}^{(r)}$ per $k=1,2,...,p-1$. Assumint independència entre observacions, el model lineal que usarem és:

$$y_i \sim Binomial\left(N_i, \pi_i = \frac{\exp\left(\beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)} + ... + \beta_{p-1}^{(r)} Ind_{p-1i}^{(r)}\right)}{1+\exp\left(\beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)} + ... + \beta_{p-1}^{(r)} Ind_{p-1i}^{(r)}\right)} \right). \qquad (6.11)$$

El nombre de punts de canvi, $p-1$, està limitat pel nombre d'observacions, n.

Per estimar $\boldsymbol{r}=(r_1, r_2,\ldots, r_{P-1})$ ajustem el model (6.11) per totes les combinacions de valors de r_k tals que $1 \le r_k \le n-1$ i $r_k > r_{k-1}$, i prenem $\boldsymbol{r}$ que millor ajusta les dades, d'acord amb un criteri de bondat de l'ajust especificat. Nosaltres triem maximitzar el màxim del logaritme de la versemblança. D'aquesta manera r és estimat com:

$$\hat{r}_{Lp} = \max_{1 \le r \le n-1} L\left(r, \hat{\beta}^{(r)}\right).$$

Si abans de començar l'anàlisi el nombre de punts de canvi no és conegut, es poden fer servir les tècniques d'obtenció de la millor equació de regressió, comprovant quines variables indicadores tenen coeficients significatius, que són les que marquen els punts de canvi existents, i quines tenen coeficients no significatius. Es pot simplificar la metodologia fixant $p_{max}-1$, el nombre màxim de punts de canvi, i obtenir estimacions de $\boldsymbol{r}=(r_1,r_2,\ldots,r_{P-1})$, per $p=2,\ldots p_{max}$, i comparar els $p_{max}-1$ models obtinguts amb les tècniques clàssiques de regressió múltiple. El millor model ens donarà l'estimació del nombre de punts de canvi i llur localització.

Aquest mètode es pot generalitzar per totes les distribucions tractades:

- Seqüències univariants de variables Normals: fent servir la regressió lineal múltiple, amb $p-1$ variables indicadores com a explicatives
- Seqüències univariants de variables binomials: fent servir la regressió per dades binàries amb $p-1$ variables indicadores com a explicatives
- Seqüències multinomials: fent servir, per la regressió politòmica amb $p-1$ variables indicadores com a explicatives, o emprant alguna de les aproximacions univariants al problema multinomial, també amb $p-1$ variables indicadores com a regressores.

En el cas de més d'un punt de canvi, la representació gràfica de l'estadístic a optimitzar (sigui la F_r per dades Normals, o el màxim del logaritme de la versemblança) en funció de r només es pot fer en el cas de 2 punts de canvi. En aquest cas, proposem representar la superfície a través de les seves corbes de nivell. La figura 6.3 mostra un exemple en el que es busquen dos punts de canvi per a l'índex de Simpson, en el que s'han ajustat models per a dades Normals. Nosaltres buscarem més d'un punt de canvi només per seqüències de dades contínues.

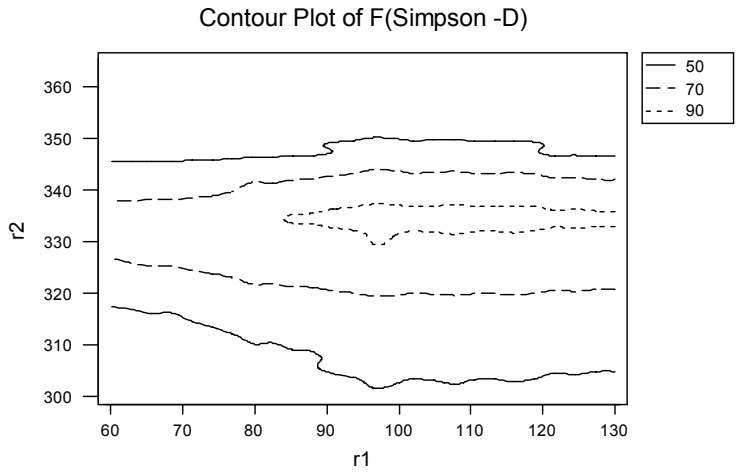

Figura 6.3: Corbes de nivell per a l'estadístic F_r de la taula ANOVA, obtinguts en ajustar els models per estimar la localització de 2 punts de canvi $\boldsymbol{r}=(r_1, r_2)$. La seqüència que s'ha analitzat és la de l'índex de Simpson, emprat per estudiar la diversitat de l'estil, en els 396 blocs de 1000 paraules consecutives en que s'ha dividit el *Tirant*.

6.6 Conclusions

Els mètodes emprats per a l'estimació de *p-1*punts de canvi, amb $p \geq 2$, basats en l'ajust de models en els que les respostes, y_i, són les observacions en la seqüència, i *p-1* variables indicadores com a explicatives, i prendre el model que millor ajusta les dades, combinat amb l'ús del gràfic que representa l'evolució de estadístics de bondat de l'ajust (log-versemblança o *F*), presenta notables avantatges respecte als mètodes basats en estimadors màxim-versemblants o en els estimadors proposats per Wolfe i Chen (1990). Els podem resumir en que:

- el mètode és senzill d'aplicar,
- el mètode gràfic permet apreciar si el punt de canvi és distingible, si hi ha ambigüitats, si hi ha altres possibles màxims que indueixin a pensar en que pot haver-hi una regió frontera en lloc d'un punt de canvi o, fins i tot, si hi ha més d'un punt de canvi,
- té una extensió natural per a detectar més de un punt de canvi,
- el mètode és vàlid per moltes distribucions diferents. Només cal trobar el model que millor ajusti la distribució de les dades,
- l'extensió del cas univariant al multivariant és molt natural i senzilla.