TÍTOL TREBALL

Artur Xarles Esparraguera

March 7, 2021

Abstract

1 Metodologia

1.1 Poisson Bivariant

Definim la distribució de Poisson Bivariant com:
Def: $(X, Y) \sim BP(\lambda_1, \lambda_2, \lambda_3)$, $\lambda_i > 0$ si $\exists$ X_1, X_2, $X3$ tal que $X_i \sim Pois(\lambda_i)$ independents i $X = X_1 + X_3$ i $Y = X_2 + X_3$.
Podem expressar la seva funció de probabilitat de la següent manera:

$$P\{X = x, Y = y\} = P\{i : (X_1, X_2, X_3) = (x - i, y - i, i)\} =$$
$$\sum_{i=0}^{min(x,y)} \frac{e^{-\lambda_1}\lambda_1^{x-i}}{(x-i)!}\frac{e^{-\lambda_2}\lambda_2^{y-i}}{(y-i)!}\frac{e^{-\lambda_3}\lambda_3^{i}}{i!} =$$
$$e^{-(\lambda_1+\lambda_2+\lambda_3)}\lambda_1^x\lambda_2^y \sum_{i=0}^{min(x,y)} \frac{1}{(x-i)!(y-i)!i!}(\frac{\lambda_3}{\lambda_1\lambda_2})^i = {}^1$$
$$e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!} \sum_{i=0}^{min(x,y)} \binom{x}{i}\binom{y}{i} i!(\frac{\lambda_3}{\lambda_1\lambda_2})^i \quad (1)$$

Aquesta distribució permet dependència entre X i Y ja que $Cov(X, Y) = Cov(X_1 + X_3, X_2 + X_3) = Cov(X_1, X_2) + Cov(X_1, X_3) + Cov(X_3, X_2) + Cov(X_3, X_3) = 0 + 0 + 0 + Var(X_3) = \lambda_3$. El paràmetre λ_3 doncs, expresa la dependència entre X i Y. De forma marginal, les variables X i Y queden representades per la suma de dues distribucions de Poisson independents, de manera que $X \sim Pois(\lambda_1 + \lambda_3)$ i $Y \sim Pois(\lambda_2 + \lambda_3)$.

[1] S'utilitza $\binom{a}{b} = \frac{a!}{b!(a-b)!}$

1.2 GLM (Model Lineal Generalitzat)

Tal com el seu nom indica, el model lineal generalitzat (GLM) és una generalització del model de regressió lineal. Així doncs, primer observem el funcionament d'aquest. Es considera una variable resposta Y de distribució Normal, un conjunt de covariables $X = (X_1, ..., X_p)$ i un vector de coeficients desconeguts associats a les covariables, $\beta = (\beta_0, \beta_1, ..., \beta_p)$. A partir de les covariables i els coeficients podem expressar el predictor lineal com $L(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$, on $L(X) \in \mathbb{R}$. El model de regressió lineal assumeix que un valor de la variable resposta Y vindrà donat a partir del predictor lineal $L(X)$ i un error, $Y = L(X) + \varepsilon$, on $\varepsilon \sim N(0, \sigma^2)$. Així doncs, $E[Y] = \mu = L(X)$.

A diferència del model de regressió lineal, en el model lineal generalitzat es pot assumir una distribució de Y diferent de la distribució Normal, com per exemple la distribució de Poisson o la de Bernoulli. En aquests casos a vegades no és correcte suposar que $E[Y] = L(X)$, com quan el domini de $E[Y]$ és tan sols un subgrup de $\mathbb{R}$. Per a solucionar aquest problema és necessari utilitzar una funció "link" que estableixi la relació entre el predictor lineal i la mitjana de la funció de distribució de Y. Establerta la funció "link", $g()$, es té $g(E[Y]) = L(X)$.

Un cop s'assumeix la distribució de Y i la funció "link" a utilitzar s'ha d'estimar el vector de paràmetres desconeguts β a partir d'una mostra de n observacions independents $(y_1, x_1), ..., (y_n, x_n)$. El mètode utilitzat per estimar β és a través de maximitzar la funció de versemblança, és a dir, trobar el conjunt de paràmetres que maximitzin:

$$\mathcal{L}(\beta|\{(y_i, x_i)\}) = \prod_{i=1}^{n} P(y_i \cap x_i)$$

Aquesta funció variarà segons la distribució que segueixi la variable resposta Y i s'ha de maximitzar a través de mètodes numèrics. En cas que la distribució de Y formi part de la família exponencial, podem garantir que els estimadors de màxima versemblança seran estadístics suficients. A més, aquests estimadors tenen la propietat asimptòtica que $\hat{\beta}_j \overset{asym.}{\sim} N(\beta_j, Var(\hat{\beta}_j))$. Un dels mètodes utilitzats per a calcular els coeficients és l'algoritme "Mínims quadrats iterativament ponderats" (IWLS). Aquest algoritme consisteix en els següents passos:

1. Obtenir una primera estimació pels paràmetres. Una opció és realitzar una regressió lineal entre la transformació de la variable resposta observada a través de la funció "link" i la resta de variables explicatives. A través del mètode de mínims quadrats ordinaris podem obtenir:

$$\hat{\beta}_j = (X^t X)^{-1}(X^t g(Y))$$

2. A partir de les estimacions, obtenir els valors predits del predictor lineal i a partir d'aquests, a través de l'inversa de la funció "link", els valors

esperats $\hat{\mu}_i$.

3. Construir una variable de resposta de treball Z_i que reflexa la diferència entre el valor esperat i el valor observat, i una matriu de ponderació W que té en la diagonal les variàncies dels valors esperats $\hat{\mu}_i$.

4. Fer una regressió de la variable Z a través de les variables explicatives X considerant les ponderacions W, de manera que podem obtenir una estimació a través de mínims quadrats ordinares:

$$\hat{\beta}_j = (X^t W X)^{-1}(X^t W Z)$$

5. Repetir els passos 2, 3 i 4 iterativament fins que convergeixi.

1.2.1 Regressió de Poisson

Com a exemple, veurem el funcionament del GLM en el cas concret de suposar que $Y \sim Pois(\lambda)$, de manera que $P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$. Es té que $E[Y] = \lambda \in (0, +\infty)$. Com que $L(X) \in \mathbb{R}$ s'ha d'utilitzar una funció "link" que passi $(0, +\infty) \mapsto (-\infty, +\infty)$. La que s'utilitza en aquest cas és el logaritme neperià, de manera que podem expressar la mitjana de la distribució a partir del predictor lineal de la següent manera:

$$ln(\lambda) = L(X) \to \lambda = e^{L(X)} \tag{2}$$

Així doncs, establerta la distribució de Y i la relació entre la seva mitjana i el predictor lineal a través de la funció "link", es procedeix a estimar els coeficients a través de maximitzar la funció de versemblança. Si tenim un conjunt d'observacions $(y_1, x_1), ..., (y_n, x_n)$ la funció és la següent:

$$\mathcal{L}(\beta|\{(y_i, x_i)\}) = \prod_{i=1}^{n} P(y_i \cap x_i) = \prod_{i=1}^{n} P(y_i|x_i)P(x_i) \propto \prod_{i=1}^{n} P(y_i|x_i) =$$
$$\prod_{i=1}^{n} \frac{(\lambda|x_i)^{y_i}}{y_i!} e^{-(\lambda|x_i)} \propto \prod_{i=1}^{n} (\lambda|x_i)^{y_i} e^{-(\lambda|x_i)} = {}^{2} \prod_{i=1}^{n} e^{L(X)y_i} e^{-e^{L(X)}} =$$
$$\prod_{i=1}^{n} e^{L(X)y_i - e^{L(X)}} = exp[\sum_{i=1}^{n} (y_i L(X) - e^{L(X)})]$$

Maximitzar aquesta funció equival a maximitzar el seu logaritme, de manera que es pot maximitzar la següent funció de log-versemblança:

$$ln(\mathcal{L}(\beta|\{(y_i, x_i)\})) = \sum_{i=1}^{n} (y_i L(X) - e^{L(X)})$$

[2] Utilitzem 2.

Com hem comentat, l'estimació dels paràmetres la podem fer a través de l'algoritme IWLS. Un cop estimats, ja s'ha ajustat el model i es pot utilitzar per a fer prediccions de λ donat el vector de covariables X i els coeficients estimats $\hat{\beta}$ a partir d'aplicar la fórmula proposada $\hat{\lambda}|X = e^{L(X)} = e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j}$.

1.2.2 Regressió de Poisson bivariant

Un cop vist el funcionament del model lineal generalitzat amb la distribució de Poisson, veurem com aplicar-ho si es té una variable resposta bivariant i s'assumeix que segueix la distribució Poisson bivariant. Per evitar confusions de notació, a partir d'aquí anomenarem al conjunt de covariables W. Així doncs, es té $(X, Y) \sim BP(\lambda_1, \lambda_2, \lambda_3)$. D'acord a la distribució proposada en la secció 1.1, $\lambda_i = E(X_i)$ per $i = 1, 2, 3$. Com que es tenen tres paràmetres corresponents a l'esperança de les corresponents distribucions de Poisson, s'han de considerar tres predictors lineals. Igual que en la regressió de Poisson la funció "link" utilitzada és el logaritme neperià i per tant tenim:

$$ln(\lambda_i) = L_i(W) \rightarrow \lambda_i = e^{L_i(W)} \tag{3}$$

per a valors de $i = 1, 2, 3$.

Igual que amb la regressió de Poisson es troba la funció de versemblança a maximitzar:

$$\mathcal{L}(\beta|\{(x_i, y_i), w_i\}) = \prod_{i=1}^{n} P((x_i, y_i) \cap w_i) = \prod_{i=1}^{n} P((x_i, y_i)|w_i)P(w_i) \propto$$

$$\prod_{i=1}^{n} P((x_i, y_i)|w_i) = \prod_{i=1}^{n} e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^{x_i}}{x_i!} \frac{\lambda_2^{y_i}}{y_i!} \sum_{j=0}^{min(x_i,y_i)} \binom{x_i}{j} \binom{y_i}{j} j! (\frac{\lambda_3}{\lambda_1\lambda_2})^j \propto$$

$$\prod_{i=1}^{n} e^{-(\lambda_1+\lambda_2+\lambda_3)} \lambda_1^{x_i} \lambda_2^{y_i} \sum_{j=0}^{min(x_i,y_i)} \binom{x_i}{j} \binom{y_i}{j} j! (\frac{\lambda_3}{\lambda_1\lambda_2})^j = {}^{3}$$

$$\prod_{i=1}^{n} e^{-(e^{L_1(W)}+e^{L_2(W)}+e^{L_3(W)})} e^{L_1(W)x_i} e^{L_2(W)y_i} \sum_{j=0}^{min(x_i,y_i)} \binom{x_i}{j} \binom{y_i}{j} j! \cdot$$

$$(\frac{e^{L_3(W)}}{e^{L_1(W)} e^{L_2(W)}})^j = \prod_{i=1}^{n} e^{-(e^{L_1(W)}+e^{L_2(W)}+e^{L_3(W)})+L_1(W)x_i+L_2(W)y_i} \cdot$$

$$\sum_{j=0}^{min(x_i,y_i)} \binom{x_i}{j} \binom{y_i}{j} j! e^{(L_3(W)-L_1(W)-L_2(W))j}$$

La corresponent log-versemblança es correspon a la següent funció:

[3] Utilitzem 3.

$$ln(\mathcal{L}(\beta|\{(X_i,Y_i),W_i\})) = \sum_{i=1}^{n} -(e^{L_1(W)} + e^{L_2(W)} + e^{L_3(W)}) + L_1(W)X_i +$$
$$L_2(W)Y_i + ln(\sum_{j=0}^{min(X_i,Y_i)} \binom{x_i}{j}\binom{y_i}{j} j! e^{(L_3(W)-L_1(W)-L_2(W))j})$$

Maximitzar aquesta funció a través de mètodes convencionals és computacionalment bastant costós, ja que conté un sumatori dins d'un altre sumatori. Així doncs, proposarem un altre mètode per a obtenir els estimadors de màxima versemblança, el mètode EM (estimació i maximització) presentat en la secció 1.3.

1.3 Algoritme EM

L'algoritme EM és un mètode a través del qual es poden obtenir aproximacions dels estimadors de màxima versemblança. És un mètode que s'utilitza principalment en presència de variables latents, és a dir, variables que no observem directament en la nostra mostra. El funcionament d'aquest algoritme es basa a anar iterant entre dos passos. El primer pas ("E-step") intenta estimar la variable latent, mentre que el segon pas ("M-step") intenta optimitzar els paràmetres del model. Un cop els paràmetres convergeixen s'atura la iteració.

Observarem el funcionament d'aquest algoritme en el nostre cas, on intentem estimar els paràmetres d'una regressió Poisson bivariant. Podem veure que tenim la presència de variables latents, ja que observem (X, Y), i la distribució té les variables latents (X_1, X_2, X_3). Tot i això, podem expressar-les a través de les dues variables observades i una sola variable latent Z:

$$X_1 = X - Z$$
$$X_2 = Y - Z$$
$$X_3 = Z$$

Un cop establerta la variable latent, s'ha de buscar els dos passos de l'algoritme EM:

1. E-step:
   En aquest pas s'ha d'estimar la variable latent Z a partir de les variables observades X i Y, i del conjunt de paràmetres $\theta = (\lambda_1, \lambda_2, \lambda_3)$ en el pas anterior (θ^*). Es fa a partir de l'esperança de la Llei $Z|X, Y, \theta^*$:

$$\begin{aligned} f(Z|X,Y,\theta^*) &= \frac{f(X,Y|Z,\theta^*)f(Z|\theta^*)}{f_{BP}(X,Y|\theta^*)} \\ &= \frac{f(X_1 = X - Z|\theta^*)f(X_2 = Y - Z|\theta^*)f(Z|\theta^*)}{f_{BP}(X,Y|\theta^*)} \end{aligned}$$

Observem que condicionar respecte θ^* simplement ens indica que tenim unes estimacions per als paràmetres λ_1, λ_2 i λ_3 en la iteració anterior. Podem calcular l'esperança de la llei per estimar el valor de Z:

$$E[Z|X,Y] = \sum_{i=0}^{min(x,y)} i\frac{f_1(x-i)f_2(y-i)f_3(i)}{f_{BP}(x,y)} =$$

$$\frac{1}{f_{BP}(x,y)} \sum_{i=1}^{min(x,y)} i\frac{e^{-\lambda_1}\lambda_1^{(x-i)}}{(x-i)!}\frac{e^{-\lambda_2}\lambda_2^{(y-i)}}{(y-i)!}\frac{e^{-\lambda_3}\lambda_3^{(i)}}{i!} =$$

$$\frac{\lambda_3}{f_{BP}(x,y)}e^{-(\lambda_1+\lambda_2+\lambda_3)}\lambda_1^{(x-1)}\lambda_2^{(y-1)} \sum_{i=1}^{min(x,y)} \frac{\lambda_1^{-(i-1)}\lambda_2^{-(i-1)}\lambda_3^{i-1}}{(x-i)!(y-i)!(i-1)!} = {}^{4}$$

$$\frac{\lambda_3}{f_{BP}(x,y)}e^{-(\lambda_1+\lambda_2+\lambda_3)}\lambda_1^{(x-1)}\lambda_2^{(y-1)} \sum_{k=0}^{min(x-1,y-1)} \frac{(\frac{\lambda_3}{\lambda_1\lambda_2})^k}{(x-k-1)!(y-k-1)!k!}$$

S'observa que $e^{-(\lambda_1+\lambda_2+\lambda_3)}\lambda_1^{(x-1)}\lambda_2^{(y-1)} \sum_{k=0}^{min(x-1,y-1)} \frac{(\frac{\lambda_3}{\lambda_1\lambda_2})^k}{(x-k-1)!(y-k-1)!k!} = f_{BP}(x-1,y-1)$, de manera que tenim:

$$E[Z|X,Y] = \lambda_3\frac{f_{BP}(x-1,y-1)}{f_{BP}(x,y)}$$

Es pot apreciar que en el cas que x o y sigui 0, $f_{BP}(x-1,y-1)$ serà 0 i per tant podem estimar el valor de la variable latent Z a partir de la següent expressió:

$$E[Z|X,Y] = \begin{cases} 0 & \text{si } x \cdot y = 0 \\ \lambda_3\frac{f_{BP}(x-1,y-1)}{f_{BP}(x,y)} & \text{altrament} \end{cases}$$

2. M-step: Un cop estimat el valor de Z i sabent els valors observats de X i Y podem trobar fàcilment els valors de X_1, X_2 i X_3. En tractar-se de tres Poissons independents d'acord amb la distribució Poisson bivariant, el pas de maximització consisteix simplement a fer una regressió de Poisson per a les tres variables. Així doncs, s'ha de calcular:

$$\hat{\beta}_1 = \hat{\beta}(x-z,W)$$
$$\hat{\beta}_2 = \hat{\beta}(y-z,W)$$
$$\hat{\beta}_3 = \hat{\beta}(z,W)$$

[4] Fem el canvi de variables: $k = i-1$

on $\hat{\beta}(x, W)$ son els paràmetres estimats per màxima versemblança d'una regressió de Poisson amb resposta x i matriu de dades W.

A partir de $\hat{\beta}_1$, $\hat{\beta}_2$ i $\hat{\beta}_3$ es pot estimar λ_1, λ_2 i λ_3, valors que necessitem en el E-step.

Hem observat per tant com trobar els dos passos de l'algoritme EM en el cas de tindre una regressió de Poisson bivariant. Així doncs, podem trobar les estimacions dels paràmetres a partir de seguir els següents passos:

1. Donar valors inicials dels paràmetres λ_1, λ_2 i λ_3 per a cada observació.
2. Realitzar els següents dos passos fins que convergeixi:
   (a) E-step: Donat el conjunt de paràmetres $\theta^* = (\lambda_1, \lambda_2, \lambda_3)$ de l'anterior iteració calcular:
   $$z_i = \begin{cases} 0 & \text{si } x \cdot y = 0 \\ \lambda_{3i} \frac{f_{BP}(x_i - 1, y_i - 1)}{f_{BP}(x_i, y_i)} & \text{altrament} \end{cases}$$
   on f_{BP} és la funció de probabilitat de la distribució Poisson bivariant.
   (b) M-step: Trobar:
   $$\hat{\beta}_1 = \hat{\beta}(x - z, W)$$
   $$\hat{\beta}_2 = \hat{\beta}(y - z, W)$$
   $$\hat{\beta}_3 = \hat{\beta}(z, W)$$
   on $\hat{\beta}$ és l'estimador de màxima versemblança d'una regressió de Poisson. Estimar λ_1, λ_2 i λ_3 a partir de $\hat{\beta}_1$, $\hat{\beta}_2$ i $\hat{\beta}_3$.
3. Un cop convergeixi ja tindrem estimats els paràmetres $\hat{\beta}_1$, $\hat{\beta}_2$ i $\hat{\beta}_3$ corresponents als coeficients dels predictors lineals $L_1(W)$, $L_2(W)$ i $L_3(W)$.

Cal destacar que podem tindre un conjunt de covariables diferent per a cada un dels paràmetres, de manera que podem utilitzar diferents variables per estimar λ_1, λ_2 i λ_3.