מעבדה לסטטיסטיקה: מודל פוואסון עם פרמטר משתנה

כזכור $\mathrm{Y_i}$ הוא מספר הפרגמנטים המתחילים בבסיס הi, i=1,...,I ואילו $j=1,\dots,J$ $\mathrm{X_j}$ משתנים מקריים המציינים לאן נפל הפרגמנט הj. בשבועיים החולפים הראינו שהנחת האחידות לא מתארת היטב את הנתונים שלנו, ואנחנו צריכים מודל גמיש יותר. מודל הפוואסון המשתנה אכן גמיש בהרבה, ומהווה את המודל הבסיסי לתופעות דיסקרטיות רבות.

מודל פוואסוני לבסיס בודד

ע״פ מודל זה, נמשיך להניח שכל הפרגמנטים מתפלגים אחיד (הנחה 1), ונזכיר את הסימון

$$p_i = P(X_1 = i) = P(X_2 = i) = \dots = P(X_J = i)$$

עבור ההסתברות שפרגמנט כלשהו יתחיל בבסיס הi (הסתברות זהה לכל פרגמנט).

כמו כן, נניח שהפרגמנטים אינם תלויים (ההנחה כנראה איננה בדיוק נכונה, אך ניתן להאמין שהתלות קטנה). בפרט, משתני האינדיקטור $1(X_j = i)$ הינם משתני ברנולי בלתי תלויים ושווי התפלגות, עם הסתברות הצלחה p_i. מכאן, ש $\mathrm{Y_i} = \sum 1(X_j = i)$ מתפלג בינומית:

$$\mathrm{Y_i} \sim Binomial(J, p_i),$$

ובקירוב פוואסוני

$$\mathrm{Y_i} \dot{\sim} Poisson(J \cdot p_i).$$

מודל פוואסוני לתא

היתרון הגדול שמאפשר הקירוב הפוואסוני זה שקל מאוד לחבר משתנים פוואסונים בלתי תלויים.

תזכורת:

א. $\mathrm{Y_1}, \mathrm{Y_2}$ משתנים פוואסונים בלתי תלויים אז $\mathrm{Y_1} + Y_2 \sim Poisson(E[Y_1] + E[Y_2])$.

ב. $\mathrm{Y_1}, \dots, \mathrm{Y_n}$ משתנים פוואסונים בלתי תלויים אז $\mathrm{Y_1} + \dots + Y_n \sim Poisson(\sum_{i=1} E[Y_i])$.

עבור תאים בגודל b, נסמן בעזרת $\tilde{Y}_{\mathrm{k}}$ את מספר הפרגמנטים בתא הk, כלומר:

$$\tilde{Y}_{\mathrm{k}} = Y_{bk-b+1} + \dots + Y_{bk}$$

אם כן, מהנחת חוסר התלות:

$$\tilde{Y}_{\mathrm{k}} \sim Poisson(\lambda_k), \qquad \lambda_k = J \cdot (p_{bk-b+1} + \dots + p_{bk}).$$

הערות:

1. בהמשך נזניח את הטילדה ב $\tilde{Y}_{\mathrm{k}}$ ונסמן במקומו $\mathrm{Y_k}$, כאשר האינדקס (k) וההקשר יבהירו שמדובר בתא ולא בבסיס יחיד.

2. כזכור, השונות של פוואסון שווה לתוחלת. כלומר $\mathrm{Var}(\mathrm{Y_k}) = E[Y_k] = \lambda_K$. בפרט, ככל שהתוחלת גבוהה יותר גם השונות גדלה. עם זאת, כדאי לזכור שהמשמעות של המשוואה היא שהפיזור (סטיית התקן, או ממוצע המרחקים) גדלה בסדר גודל של שורש התוחלת.

3. <u>פירוק השונות</u>: כאשר מסתכלים על השונות הנצפית בין ערכי $\mathrm{Y_k}$ השונים, נשים לב שישנם שני מרכיבים לשונות. ראשית, התוחלות של כל Y_k אחרת, ושנית, כל Y_k מתפלג סביב תוחלתו.למעשה, השונות הנצפית היא סכום של שני רכיבים אלו:

$$\mathrm{Var}[\mathrm{Y_k}] = \mathrm{E}_{\mathrm{k}\in K}\left[\mathrm{Var}_{\mathrm{Pois}}[\mathrm{Y_k}]\right] + \mathrm{Var}_{\mathrm{k\in K}}\left[\mathrm{E}_{\mathrm{Pois}}[\mathrm{Y_k}]\right]$$
$$= \mathrm{E}_{\mathrm{k}\in K}[\lambda_\mathrm{k}] + \mathrm{Var}_{\mathrm{k\in K}}[\lambda_\mathrm{k}] = \bar{\lambda} + \frac{1}{K}(\lambda_k - \overline{\bar{\lambda})}^2.$$

כאשר $\bar{\lambda} = \frac{1}{K}\sum\lambda_k = J \cdot b/I$ היא התוחלת הממוצעת לתא בגודל b.

4. אפילו אם $p_1, \dots, p_I$ שונים אחד מהשני, זה לא מספיק כדי לראות הבדלים גדולים בין הλ_kות. הסיבה להבדלים הגדולים שאנחנו בין הλ_k היא שההסתברויות בסביבה הקרובה דומות אחת לשניה, ולכן לא מתבטלות כאשר נסכמים יחדיו. אם לא היה מבנה מרחבי ובאופן מקרי היינו מאחדים תאים, רוב ההבדלים היו נעלמים.

5. לפחות בשלב זה ננסה לחזות את λ_k ישירות ע״י התאמת קו רגרסיה מ$\mathrm{GC_k}$, אולם המודל מזמין גם לנסות לחזות את ה $\mathrm{p_i}$ים ישירות. אני מקווה שנראה עוד מזה בשבועות האחרונים של המעבדה.