תסריט להקלטה למחברת 2: בייסאני מול OLS/MixedLM/ANOVA, טיפים במסעדה

משך משוער: כ-29-34 דקות בהקלטה בפועל (קריאה בקצב בינוני, ללא זמן המתנה על ריצת תאים).

פתיחה

במחברת הזאת אני משווה רגרסיה, פעם בגישה הבייסאנית ופעם בגישה הקלאסית, כדי לבדוק שאלה אחת: האם סכום החשבון במסעדה מנבא את גובה הטיפ, והאם הקשר הזה שונה בין ימות השבוע. הדאטה הוא מאגר אמיתי של 244 חשבונות מסעדה, מובנה בתוך ספריית seaborn. אני משתמשת בכמה מקבילות קלאסיות שונות לאורך המחברת: מול המודל הפשוט אני משווה לרגרסיית OLS רגילה, מול השאלה של הבדלים בין ימים אני משווה גם למבחני ANOVA וגם למודל אפקטים מעורבים שנקרא MixedLM, שהוא הגרסה הקלאסית למודל היררכי.

לצד הספריות הרגילות, נטען כאן גם המודול שמאפשר להריץ MixedLM.

הדאטה, ופרדוקס סימפסון

לפני שממשיכים לבניית מודלים, בואו נסתכל רגע על הדאטה עצמו. עוד לפני הגרפים, יש טבלת describe מפורקת לפי יום: לכל אחד מארבעת הימים כמות התצפיות, ממוצע, סטיית תקן, מינימום, מקסימום ואחוזונים, גם של סכום החשבון וגם של הטיפ. כבר כאן רואים נקודה שתחזור לאורך כל המחברת: ליום שישי יש רק תשע-עשרה תצפיות, הרבה פחות מהימים האחרים (בין שישים לשמונים-ושבע), מה שיהפוך אותו לקבוצה הכי רועשת ולזאת שכיווץ ישפיע עליה הכי הרבה בהמשך.

גרף הפיזור הזה: סכום החשבון על ציר ה-X, גובה הטיפ על ציר ה-Y, כל יום בשבוע בצבע אחר עם קו רגרסיה משלו, ומעליהם קו שחור מקווקו שמייצג רגרסיה אחת מאוחדת על כל הימים ביחד. הרעיון הוא לבדוק חזותית אם קווי הרגרסיה הצבעוניים, כל יום בנפרד, הולכים לכיוון שונה מהקו השחור המאוחד. שאלתי את הצ׳אט על תופעה שהתמונה הזאת מזכירה:

> **Chat, what is Simpson's paradox in one sentence, and how do you deal with it in another? Does it show up in this restaurant-tips data?**

והוא ענה שזה מצב שבו מגמה שמתקיימת בתוך כל תת-קבוצה בנפרד יכולה להתהפך או להיעלם לגמרי ברגע שמאחדים את כל התת-הקבוצות יחד, והדרך להתמודד איתה היא תמיד לבדוק את הקשר בתוך הקבוצות הטבעיות לפני שסומכים על הגרסה המאוחדת. זה בדיוק מה שמצדיק בהמשך לבנות מודל שמתייחס לכל יום בנפרד, במקום רגרסיה אחת גורפת.

לא הסתפקתי בהגדרה המילולית ובדקתי אם זה בכלל קיים בדאטה שלי: חישבתי את השיפוע של קו הרגרסיה בנפרד לכל יום, ואת השיפוע של הרגרסיה המאוחדת. כל ארבעת הימים יצאו עם שיפוע חיובי, וגם השיפוע המאוחד יצא חיובי, כך שאין כאן היפוך סימן מלא, וזו לא דוגמה טהורה לפרדוקס סימפסון כמו המקרים הקלאסיים שבהם מגמה שלמה מתהפכת. מה שכן יש הוא הבדל אמיתי בגודל השיפוע בין הימים, מתחת ל-0.07 ביום ראשון ועד מעל 0.12 בחמישי, וזו גרסה מתונה יותר של אותה בעיה בדיוק: רגרסיה מאוחדת אחת מטשטשת התנהגות שבפועל שונה בין ימים.

המודל שאני בודקת כאן פשוט: הטיפ שווה חיתוך, ועוד שיפוע כפול סכום החשבון, ועוד רעש נורמלי אקראי. אפשר לכתוב את זה בשתי צורות שקולות לגמרי: הטיפ שווה לחיתוך ועוד שיפוע כפול סכום החשבון, ועוד רעש נורמלי בממוצע אפס; או, באופן שקול, הטיפ עצמו מתפלג נורמלית סביב חיתוך ועוד שיפוע כפול סכום החשבון. הצורה השנייה היא זו שממופה ישירות לקוד ה-PyMC בעוד רגע, שם כותבים ישירות "הטיפ מתפלג נורמלית עם ממוצע כזה וסטיית תקן כזאת". קודם, הגרסה הקלאסית: רגרסיית OLS. שאלתי את הצ'אט מה חשוב לבדוק בפלט הזה:

> **Chat, what should I look for in this statsmodels OLS output, and what are OLS's assumptions?**

והוא ענה שקודם מסתכלים על מקדם סכום החשבון, שאומר כמה כל דולר נוסף בחשבון מוסיף בממוצע לטיפ, עם ערך מובהקות זעיר שאומר שהקשר בטוח לא אפס במקרה. ה-R-squared אומר כמה אחוז מהשונות בטיפ מוסבר על ידי המשתנה הזה לבדו. הנחות OLS: קשר לינארי, שגיאות בלתי-תלויות, שונות קבועה של השגיאות לאורך כל טווח הערכים, ונורמליות של השגיאות, שחשובה בעיקר למדגמים קטנים כדי שהמבחנים הסטטיסטיים יהיו מדויקים.

הפלט הוא טבלה שלמה, וכדאי לעבור עליה שורה-שורה כי היא תחזור שוב בהמשך הקורס. ערך ה--R squared יצא כמעט 46 אחוז, מה שאומר שסכום החשבון לבדו מסביר קרוב למחצית מהשינוי בטיפ, לא הכל, הגיוני, כי טיפ תלוי גם בשירות, בגודל הקבוצה ובמצב הרוח. סטטיסטי F, יחד עם ההסתברות הזעירה שלו, בודק האם המודל כולו משפר את החיזוי בצורה מובהקת לעומת לא להשתמש בשום משתנה, וכאן התשובה ברורה וחד-משמעית.

בשורה של סכום החשבון בטבלת המקדמים יש כמה מספרים שחשוב להבין: המקדם עצמו, קרוב ל-0.105, אומר שכל דולר נוסף בחשבון מוסיף בערך עשרה וחצי סנט לטיפ הצפוי. השגיאה התקנית שלצדו מבטאת את אי-הוודאות סביב האומדן הזה. סטטיסטי ה-t הוא היחס בין השניים, כמה סטיות תקן האומדן רחוק מאפס, ומספר גדול כמו זה שיצא כאן אומר שהקשר בטוח לא אפס במקרה. לצדו יש את ערך המובהקות של אותו מבחן, שיצא זעיר, ולבסוף שני העמודות האחרונות הן רווח הסמך של 95 אחוז למקדם עצמו, שיצא בין כתשעה לשתים-עשרה סנט בערך. אותו סוג ניתוח קיים גם לחיתוך.

לפני שאני בכלל בודקת את הנחות OLS בפועל, שאלתי את הצ'אט למה לשים לב בגרפי השאריות:

> **Chat, what should I look for in residual plots, and are the residuals here actually normal?**

והוא ענה שפיזור שאריות מול ערכים מנובאים אמור להיראות כמו ענן חסר צורה סביב האפס, וכל עקמומיות נראית לעין, צורת משפך, או מגמה, מסמנת הפרת הנחה. לבדיקת נורמליות ספציפית, תרשים Q-Q משווה את האחוזונים של השאריות לאחוזונים שהיו מתקבלים מהתפלגות נורמלית מושלמת: נקודות שנצמדות לקו האלכסוני אומרות שהשאריות נראות נורמליות, ועקמומיות שיטתית ממנו, בעיקר בקצוות, אומרת שהן לא.

כדאי לבדוק את זה בפועל ולא רק להניח שההנחות מתקיימות. שלושה גרפים זה לצד זה: הראשון הוא שאריות המודל מול הערכים המנובאים. השני הוא היסטוגרמה של השאריות. השלישי הוא תרשים ה--Q Q שכבר תואר למעלה. הרצתי גם מבחן שפירו-וילק פורמלי לנורמליות, ויצא עם ערך מובהקות זעיר, אז שאלתי את הצ׳אט מה זה בעצם אומר כאן, לאור זה שהמקדם עצמו נראה תקין:

> **Chat, what does a low Shapiro-Wilk p-value here actually mean, given the coefficient estimate itself still looked fine above?**

והוא ענה שערך מובהקות זעיר במבחן שפירו-וילק דוחה את ההנחה שהשאריות מתפלגות נורמלית, ואכן, הענן בגרף הראשון מראה נטייה קלה ימינה, והתרשים השלישי מתעקל בקצה העליון שלו, אז זו הפרה אמיתית וניתנת לזיהוי, לא אזעקת שווא. אבל זה בעיקר מאיים על הדיוק המדויק של השגיאות התקניות ורמות המובהקות במדגמים קטנים, ולא על האומדן עצמו, שנשאר חסר הטיה כל עוד שאר ההנחות מתקיימות. עם 244 תצפיות, משפט הגבול המרכזי גם עושה חלק מהעבודה שהנחת הנורמליות הייתה אמורה לעשות עבור המבחנים הסטטיסטיים.

עכשיו לאותו מודל בדיוק, בגישה הבייסאנית, ב-PyMC, וצריך לבחור פריורים. שאלתי את הצ׳אט למה הבחירות שלי סבירות:

> **Chat, why are `HalfCauchy(beta=2)` for sigma and `Normal(0, 5)` for the coefficients reasonable priors here?**

והוא ענה שסיגמא, שהיא סטיית התקן של הרעש, חייבת להיות חיובית, ולכן אי אפשר לשים עליה פריור נורמלי רגיל שנותן גם ערכים שליליים. HalfCauchy היא התפלגות עם זנב כבד מאוד, שכמעט לא מגבילה ערכים גדולים ונותנת לנתונים לדבר בעצמם, ולכן היא הבחירה הסטנדרטית לפרמטרי שונות. פריור נורמלי רחב לחיתוך ולשיפוע בעצם אומר ״הייתי מופתעת מקשר שגדול מעשרה דולר בערך״, מגבלה רחבה מאוד ביחס לטווח האמיתי של טיפים, שנע בערך בין דולר לעשרה דולר, כך שהיא כמעט לא מכתיבה תשובה ורק חוסמת ערכים אבסורדיים.

לפני שדוגמים בפועל, יש כאן דיאגרמה גרפית של מבנה המודל: סיגמא, החיתוך והשיפוע מופיעים כעיגולים פתוחים, כי הם נסתרים ולא ידועים, וכולם מזינים חץ לתוך הטיפ, שמופיע כעיגול מוצל כי הוא נצפה, יש לנו את הנתונים בפועל.

השוואה ישירה בין OLS לבייסאני

עכשיו אפשר להשוות ישירות בין שתי הגישות. הפוסטריורי מוצג בשלושה פאנלים נפרדים, אחד לכל פרמטר, חיתוך, שיפוע וסיגמא: בכל פאנל ציר ה-X הוא הערך האפשרי של הפרמטר וציר ה-Y הוא צפיפות ההסתברות, עם עקומה שמראה כמה סביר כל ערך, קו שמסמן את אומדן הנקודה, ותחום מוצלל שמסמן את רווח האמון. עוד נקודת גרסה שכדאי לציין בקול, כי בדיוק כזאת בעיה שקטה קל לפספס: בגרסת arviz הנוכחית, כשמבקשים טבלת סיכום עם רווח מסוים בלי לציין איזה סוג רווח, ברירת המחדל השתנתה לרווח דו-קצוותי, לא ל-HDI כמו שהיה בעבר. מכיוון שבגרף הבא אני קוראת לרווח הזה ״HDI״ בפירוש, ביקשתי את ה-HDI במפורש בקוד ולא הסתמכתי על ברירת המחדל השקטה, בדיוק כדי שהתווית בגרף תהיה נכונה ולא רק כמעט נכונה.

עכשיו הרגע המרכזי: המקדמים והרווחים מוצגים אחד ליד השני, בטבלה ובגרף עם שתי שורות, חיתוך ושיפוע, וכל שורה עם שתי נקודות וקווי שגיאה: נקודה אחת ל-OLS ונקודה שנייה, מתחתיה, לבייסאני. עם כמעט 244 נקודות ופריור כמעט שטוח, השיפוע הבייסאני והשיפוע של OLS יוצאים כמעט זהים ממש, וכך גם רווחי הסמך שלהם, בדיוק כמו שהתאוריה הא-סימפטוטית מנבאת. הערך המוסף של הגישה הבייסאנית כאן הוא לא תשובה שונה למודל הפשוט הזה, אלא הכלים שנבנים סביב הפוסטריורי, שנראה עכשיו.

מודל חיזוי: רצועות אי-ודאות

יש כאן גרף חדש שממחיש איך נראית תחזית מלאה: ציר ה-X הוא סכום החשבון וציר ה-Y הוא הטיפ, עם שלוש שכבות: קו ברור אחד שמייצג את הממוצע הפוסטריורי של קו הרגרסיה, רצועה צרה יחסית סביבו שמייצגת רווח אמון על מיקום הקו עצמו, ורצועה הרבה יותר רחבה שמייצגת רווח אמון על תחזית לתצפית חדשה, שכוללת גם את הרעש הרגיל. מתחת לכול, נקודות הפיזור של הדאטה האמיתי. שאלתי את הצ׳אט למה הרצועות מתנהגות אחרת זו מזו:

> **Chat, why do the predictive-interval bands stay roughly constant width while the HDI bands narrow toward the middle of the data?**

והוא ענה שהרצועה הפנימית משקפת רק חוסר ודאות על מיקום הקו, שהכי קטן במרכז הדאטה וגדל בקצוות, בצורת שעון חול. הרצועה החיצונית מוסיפה על זה גם את הרעש השיורי, שההנחה שלנו עליו קבועה בכל טווח, וזה הרבה יותר גדול מהשינוי הקטן ברוחב הרצועה הפנימית, כך שהוא בולע אותו כמעט לגמרי ונראה כמעט קבוע.

מה קורה מתחת למכסה המנוע: MCMC ושיטת הרשת

כדי להראות בפועל מה אלגוריתם ה-MCMC עצמו עושה מאחורי הקלעים, ולא רק לדבר על זה תיאורטית, יש כאן גרסה פשוטה יותר של המודל: אותו חיתוך ושיפוע, אבל הפעם עם סיגמא קבועה על אחת ולא פרמטר נלמד, כלומר הטיפ שווה לחיתוך ועוד שיפוע כפול סכום החשבון ועוד רעש נורמלי עם סטיית תקן קבועה של אחת. הסיבה לפישוט: עם רק שני פרמטרים חופשיים, חיתוך ושיפוע, אפשר להסתכל ישירות על כל הפוסטריורי המשותף שלהם בשני ממדים על גרף אחד, בלי מימד שלישי של אי-ודאות שהיה מסבך את התמונה. גרף הפיזור הדו-ממדי של הדגימות: ציר ה-X הוא ערכי החיתוך שנדגמו, ציר ה-Y הוא ערכי השיפוע שנדגמו, וכל נקודה היא דגימה אחת מהשרשרת. הענן יוצא בצורת אליפסה. שאלתי את הצ׳אט למה זה קורה, ואם אפשר לראות את זה בלי MCMC בכלל:

> **Chat, why does the joint posterior of intercept and slope look like a 2D Gaussian, and is there a way to see this without MCMC?**

והוא ענה שעם לייקליהוד נורמלי ופריורים נורמליים, הלוגריתם של הפוסטריורי הוא פונקציה ריבועית של הפרמטרים, וזו בדיוק הצורה המתמטית של גאוסיאן רב-משתני, כך שאין צריך MCMC בכלל כדי לדעת מראש איך התמונה תיראה. בשביל מודל כל כך קטן, עם שני פרמטרים בלבד, אפשר גם להשתמש בשיטת הרשת: מחשבים פריור כפול לייקליהוד על רשת דו-ממדית עדינה של ערכים אפשריים, ומנרמלים, ומקבלים את אותה תמונה בלי שום דגימה בפועל.

גרף הקונטור הזה, מפת גבהים: ציר ה-X הוא ערכי החיתוך על הרשת, ציר ה-Y הוא ערכי השיפוע, וכל קו מתאר מחבר נקודות באותה צפיפות פוסטריורית, בדיוק אותה צורה אליפטית שהתקבלה מה-MCMC למעלה, רק שהפעם היא חושבה ישירות בלי שום דגימה. שאלתי את הצ׳אט על נקודה טכנית בקוד עצמו, ולמה שיטת הרשת מוגבלת:

> **Chat, why is computing the likelihood as a raw product (rather than a sum of logs) problematic here, and why is the grid method only viable for 1-3 parameters?**

והוא ענה שעם 244 נקודות, הכפלה ישירה של הסתברויות גולמיות מתאפסת מספרית כמעט מיד, ולכן צריך לעבוד עם לוגריתם ההסתברות ולחבר במקום להכפיל, ורק בסוף להעלות בחזרה בחזקת e. שיטת הרשת עובדת רק עד שניים-שלושה פרמטרים לכל היותר, כי העלות שלה גדלה אקספוננציאלית עם מספר הפרמטרים, היא לא יודעת לאן לכוון מראש בלי לדעת איפה הפוסטריורי בכלל נמצא, וצריך לנחש מראש את טווח הערכים עוד לפני שרואים את הדאטה.

רגרסיה היררכית: MixedLM מול בייסאני

עכשיו לשאלה המרכזית של המחברת הזו: האם הקשר בין חשבון לטיפ שונה בין הימים, במיוחד ביום שישי, שיש לו רק תשע-עשרה תצפיות, הרבה פחות מהימים האחרים. בודקת את זה קודם בדרך הקלאסית, עם מודל אפקטים מעורבים. שאלתי את הצ׳אט מה זה בעצם, ואיך הוא קשור למודל ההיררכי הבייסאני:

> **Chat, what is a linear mixed-effects model, and how is it the classical analogue of a Bayesian hierarchical model?**

והוא ענה שמפרידים בין אפקט קבוע, שהוא החיתוך והשיפוע הממוצעים על פני כל הימים יחד, לבין אפקטים אקראיים, שהם כמה כל יום ספציפי סוטה מהממוצע הזה, ומניחים שהאפקטים האקראיים מגיעים מהתפלגות נורמלית משותפת שהשונות שלה נאמדת מהנתונים עצמם. זו בדיוק אותה מחשבה כמו הפריור-העל המשותף שנבנה בהמשך בגרסה הבייסאנית.

בפלט של המודל הזה יש כמה שורות שכדאי לפרק. מספר הקבוצות הוא ארבע, כמספר הימים. שיטת האמידה שכתובה שם היא נראות מוגבלת, שהיא וריאציה על נראות מרבית המיועדת במיוחד לאמידת רכיבי שונות: נראות מרבית רגילה נוטה להעריך שונות בין-קבוצתית בחסר כשיש מעט קבוצות, והשיטה הזו מתקנת את ההטיה הזאת. כשהרצתי את זה קפצה אזהרת אי-התכנסות אמיתית, וכדאי לספר עליה בקול ולא להסתיר: עם רק ארבע קבוצות, לאלגוריתם יש מעט מאוד מידע כדי לקבוע את מבנה השונות ביניהן, וזה מוכר. השתמשתי בתוצאה בכל זאת, כי ניסיון חוזר עם שיטת אופטימיזציה אחרת כן התכנס לתשובה סבירה.

בטבלת המקדמים עצמה, מקדם החיתוך יצא סביב 0.95 עם מובהקות גבוהה, ומקדם סכום החשבון יצא סביב 0.104, קרוב מאוד למה שקיבלנו ב-OLS הרגיל. השונות בין הימים עצמם, שמופיעה בשורה נפרדת, נותנת מושג כמה הימים שונים זה מזה. האומדן הנקודתי לכל יום ספציפי נקרא הטוב ביותר בין האומדנים חסרי ההטיה, ובקיצור באנגלית BLUP. בטבלה שלו: יום שישי יוצא עם סטייה קטנה מהממוצע, ואילו יום ראשון יוצא עם הסטייה הגדולה ביותר.

עכשיו הגרסה הבייסאנית, בפרמטריזציה ממורכזת: לכל יום יש חיתוך ושיפוע משלו, שנשלפים מהתפלגות נורמלית משותפת שהפרמטרים שלה הם עצמם פריורים שנאמדים מהנתונים. שוב יש כאן דיאגרמה, וההבדל מהמודל הבסיסי בולט מיד: יש שכבה נוספת של עיגולים, פריור-העל, שמעל החיתוך והשיפוע של כל יום. בקוד עצמו יש נקודה שכדאי להזכיר: אני ממרכזת את משתנה סכום החשבון לפני בניית המודל, כלומר מחסירה ממנו את הממוצע שלו. הסיבה: אם משתמשים בסכום החשבון הגולמי, שנע בין שלושה לחמישים דולר בערך, רחוק מאוד מאפס, החיתוך של המודל בעצם מייצג אקסטרפולציה ל״חשבון של אפס דולר״, ערך שכמעט לא קיים במציאות, וזה יוצר קורלציה חזקה ומלאכותית בין החיתוך לשיפוע, שפוגעת בגיאומטריה של הדגימה. אחרי מירכוז, החיתוך מייצג ״טיפ בחשבון בגודל ממוצע״, הרבה יותר יציב ופרשני.

אחרי הדגימה, גרף מסלול ספציפית ליום שישי, הקבוצה הכי קטנה: פאנל נפרד לכל פרמטר, ציר ה-X מספר הדגימה, ציר ה-Y הערך שנדגם, וארבעת הצבעים הם ארבע השרשראות, שצריכות להיראות כמו רעש חסר מבנה שחופף כמעט לגמרי. שאלתי את הצ׳אט אם צריך לדאוג משהו לגבי הקבוצה הכי קטנה הזאת:

> **Chat, should I be worried about anything in the trace/summary for the smallest group (Friday, n=19)?**

והוא ענה שקבוצה קטנה כמו יום שישי נשענת חזק על הפריורים המשותפים, ולכן שווה לבדוק ספציפית את מדדי ההתכנסות שלה: r-hat גבוה מדי או גודל מדגם אפקטיבי נמוך מדי, שני הסימנים הקלאסיים לקבוצה שקשה לאמוד באופן עצמאי.

בטבלת הסיכום רואים שהרווח סביב השיפוע של יום שישי אכן רחב יותר מזה של הימים האחרים, סימן שאכן יש פחות ודאות סביבו.

דיברגנסים: משפך של Neal

עוד כלי אבחון חשוב: גרף פיזור דו-ממדי שבו ציר ה-X הוא השיפוע ביום שישי וציר ה-Y הוא פרמטר השונות המשותף, ובו נקודות שסומנו כדיברגנסים, כלומר כישלונות של האלגוריתם, מודגשות בצבע. הן מתרכזות דווקא באזור שבו פרמטר השונות קטן מאוד. שאלתי את הצ׳אט מה אני בעצם רואה כאן, ומה גורם לזה:

> **Chat, what am I looking for in this pair plot, and what causes it (Neal's funnel)?**

והוא ענה שכשלקבוצה יש מעט מאוד מידע משלה, השיפוע שלה והפרמטר המשותף נעשים תלויים חזק זה בזה: פרמטר שונות קטן כופה את השיפוע להיות קרוב מאוד לממוצע, וצורת הפוסטריורי המקומית משתנה בצורה דרסטית בין הצוואר הרחב לגרון הצר של המשפך. גודל צעד קבוע אחד, שהאלגוריתם קובע פעם אחת בתחילת הריצה, פשוט לא יכול להתאים לשני האזורים גם יחד.

עוד שני כלי אבחון: אוטוקורלציה ו-ESS

שני כלי אבחון נוספים: גרף אוטוקורלציה על פרמטר השונות המשותף, ואחריו גרף גודל מדגם אפקטיבי בגרסה מקומית. שאלתי את הצ׳אט איך בדיוק מחשבים את זה, ולמה יש סף מומלץ של ארבע מאות:

> **Chat, how does arviz calculate the autocorrelation shown in `az.plot_autocorr`, and why is 400 a common rule-of-thumb target for ESS?**

והוא ענה שאוטוקורלציה בפיגור מסוים מודדת כמה דגימה בשרשרת מתואמת עם דגימה שהייתה באותו מרחק פיגור קודם לכן, והציפייה היא שהמתאם יורד מהר לכיוון אפס ככל שהפיגור גדל. גודל מדגם אפקטיבי מודד כמה דגימות בלתי-תלויות שוות ערך יש לנו בפועל, לא כמה דגימות יש על הנייר, כי דגימות עוקבות ב-MCMC תלויות זו בזו מעצם בנייתן. ארבע מאות היא בערך הנקודה שבה שגיאת התקן של מונטה-קרלו על אומדנים סטנדרטיים, כמו ממוצע או אחוזונים, נעשית יציבה מספיק כדי לסמוך עליה.

בגרף האוטוקורלציה ציר X הוא הפיגור וציר Y הוא מידת המתאם, ובגרף ה-ESS המקומי ציר X הוא טווח האחוזונים האפשריים של הפרמטר וציר Y הוא גודל המדגם האפקטיבי הרלוונטי לכל תחום כזה, עם קו מקווקו שמסמן את הסף המומלץ.

הפתרון: פרמטריזציה לא-ממורכזת

הפתרון למשפך הוא לשנות איך מייצגים את הפרמטר. שאלתי את הצ׳אט מה בדיוק ההבדל מהמודל הממורכז, ואיפה ראיתי טריק דומה בעבר:

> **Chat, what's the difference between this model and the centered one, and where have I seen this "offset then scale" trick before?**

והוא ענה שבמקום להצהיר על השיפוע של כל יום ישירות כנשלף מהתפלגות משותפת, מצהירים על משתנה סטנדרטי בסקאלה קבועה, ובונים את השיפוע בפועל כפעולה דטרמיניסטית: ממוצע משותף, ועוד המשתנה הסטנדרטי כפול סטיית התקן המשותפת. זו בדיוק אותה טריקה שמשמשת ב-Variational Autoencoders בלמידה עמוקה, אותה אלגברה בדיוק, רק ששם המטרה היא לאפשר גרדיאנטים לזרום דרך פעולת דגימה, וכאן המטרה היא לפשט את הגיאומטריה שהאלגוריתם צריך לחקור.

רגע כנות שחשוב לספר עליו, כי הוא בדיוק הדגמה של לבדוק תשובה ולא לקחת אותה כמובנת מאליה: הריצה הראשונה שלי, לפני מירכוז ה-X, הראתה בכלל לא שיפור דרמטי, כמעט אותו מספר דיברגנסים בשני המודלים. זה היה אמור להדאיג, כי ההסבר של משפך Neal מנבא שיפור גדול. חקרתי את זה במקום להתעלם, ומצאתי בעיה נפרדת ומקבילה: לא מירכזתי את משתנה ה-X עצמו, מה שיוצר בדיוק את בעיית הקורלציה המלאכותית בין החיתוך לשיפוע שהזכרתי קודם, והיא פוגעת בשני המודלים באותה מידה, ובכך מסתירה את התועלת האמיתית של תיקון המשפך. אחרי שמירכזתי את ה-X, קיבלתי את התוצאה האמיתית: מהמודל הממורכז לבדו, עם כ-463 דיברגנסים, למודל הלא-ממורכז, עם 10 בלבד, שיפור דרמטי ואמיתי, בדיוק כמו שהתאוריה חוזה.

כיווץ בשלוש רמות

עכשיו לבדוק את הכיווץ משלוש זוויות בבת אחת. השאלה האחרונה היא אם משווים שיפוע נפרד לכל יום, בלי שום שיתוף מידע ביניהם, מול האומדן של מודל האפקטים המעורבים, מול הממוצע הפוסטריורי הבייסאני, האם שלושתם מסכימים על אילו ימים נמשכים הכי חזק לכיוון הממוצע. שאלתי את הצ׳אט את זה במפורש:

> **Chat, compare the day-level slope estimates from: (a) fitting each day completely separately with its own OLS, (b) MixedLM's random-effect (BLUP) adjustment, and (c) the Bayesian hierarchical posterior mean. Do they agree on which days get pulled the most?**

והוא ענה שאכן כן, ובאותו כיוון: יום שישי, הקבוצה הכי קטנה, אמור לזוז הכי הרבה מהאומדן הנפרד לכיוון הממוצע המשותף גם ב-BLUP של MixedLM וגם בממוצע הפוסטריורי הבייסאני, כי שתי השיטות מיישמות את אותו רעיון של איחוד חלקי, פשוט דרך מנגנון סטטיסטי שונה: אחת ממקסמת נראות והשנייה דוגמת פוסטריורי.

הגרף כאן, עם שלושה טורים, ללא שיתוף, אפקטים מעורבים, ובייסאני, וקו לכל יום שמחבר בין שלוש הנקודות שלו בכל טור, מאמת את זה חזותית: יום שישי אכן זז הכי הרבה משני הטורים האחרים לכיוון הממוצע המשותף, בשתי השיטות. זה בדיוק אימות מספרי לטענה שמודל אפקטים מעורבים הוא המקבילה הקלאסית לכיווץ בייסאני, לא רק אנלוגיה מילולית שנטענת בלי בדיקה.

שאלה פשוטה יותר: משתי ימים לארבעה

כל מה שראינו עד עכשיו משווה את הקשר בין חשבון לטיפ, כלומר את השיפוע, בין הימים. שאלה פשוטה יותר, וקלאסית יותר, שמתעלמת מגודל החשבון לגמרי: האם אנשים פשוט משאירים סכומי טיפ שונים בימים שונים? נקודת המוצא הטבעית, לפני שמשווים את כל ארבעת הימים ביחד, היא המקרה הפשוט של שתי קבוצות: האם סכומי הטיפ הגולמיים בשבת ובשישי שונים זה מזה?

שאלתי את הצ'אט מה צריך לבדוק לפני שמריצים מבחן t, ומה האלטרנטיבה הלא-פרמטרית לו אם ההנחות לא מתקיימות:

> **Chat, what do I need to check before running a t-test, and what's the non-parametric alternative?**

והוא ענה שלפני מבחן t בלתי-תלוי לשתי קבוצות, כדאי לבדוק שכל קבוצה קרובה יחסית להתפלגות נורמלית, ושהשונויות של שתי הקבוצות דומות זו לזו. יש גרסה מתוקנת למקרה של שונויות לא שוות, שנחשבת ברירת מחדל בטוחה יותר כשלא בטוחים. כשאלה נראות בעייתיות, כפי שכבר ראינו קודם בבדיקת הנורמליות של השאריות, האלטרנטיבה הלא-פרמטרית היא מבחן מן-וויטני, שמשווה בין שתי הקבוצות רק לפי הדירוגים של הנתונים המאוחדים ולא לפי הערכים הגולמיים.

הרצתי את שני המבחנים על סכומי הטיפ בשבת ובשישי, ושניהם הגיעו לאותה מסקנה איכותית: אין הבדל מובהק בין שני הימים האלה בטיפ הגולמי. זה מרגיע, אבל זה לא תמיד יהיה כך, וההשוואה הרחבה יותר בהמשך היא בדיוק המקום שבו כבר ראינו את התשובה הפרמטרית והלא-פרמטרית נפרדות זו מזו.

עכשיו לכל ארבעת הימים ביחד. שאלתי את הצ'אט מה השערת האפס של ANOVA, ולמה זה בעצם "רק" מבחן F:

> **Chat, what's `H0` of a one-way ANOVA, what are its assumptions, and why is it "just" an F-test?**

והוא ענה שהשערת האפס היא שלכל הקבוצות אותו ממוצע אוכלוסייתי. ANOVA עובד על ידי השוואת השונות בין ממוצעי הקבוצות לשונות בתוך הקבוצות, והיחס הזה מתפלג לפי התפלגות F תחת השערת האפס, כך ש-ANOVA ומבחן F חד-כיווני הם אותו הליך בדיוק תחת שני שמות שונים. קורסים שלמים קיימים על ANOVA בעיקר בגלל איך זה מוכלל לכמה גורמים ואינטראקציות, לא כי המקרה החד-כיווני עמוק יותר ממבחן F. ההנחות שלו דומות להנחות של OLS: תצפיות בלתי-תלויות, שאריות מתפלגות נורמלית בתוך כל קבוצה, ושונות שווה בין הקבוצות, ומבחן בארטלט הוא הבדיקה הקלאסית הסטנדרטית לזה האחרון.

הרצתי את מבחן בארטלט על סכומי הטיפ בארבעת הימים, ותוצאתו הייתה מובהקת: השונויות באמת לא שוות: יום שבת מפוזר משמעותית יותר מיום שישי. זו בדיוק ההנחה ש-ANOVA נשען עליה, וכאן היא מופרת לא רק בתיאוריה אלא בנתונים עצמם. הרצתי בכל זאת את ה-ANOVA עצמו, ויצא לא מובהק. שאלתי את הצ׳אט מתי בכלל עדיף להעדיף את המבחן הלא-פרמטרי המקביל:

> **Chat, when should I prefer Kruskal-Wallis over ANOVA, and does it matter here?**

והוא ענה שקרוסקל-וואליס בודק אם המדגמים מגיעים מאותה התפלגות באמצעות הדירוגים בלבד של הנתונים המאוחדים, ולא הערכים הגולמיים, ולכן הוא לא דורש נורמליות והרבה פחות רגיש להפרת שוויון השונויות. זה הגיבוי הטבעי בדיוק כשההנחות של ANOVA נראות רופפות, כמו שראינו במבחן בארטלט. וכן, זה משנה כאן בפועל: ANOVA חוזר לא מובהק, בעוד קרוסקל-וואליס על אותם נתונים בדיוק חוזר מובהק.

אי ההתאמה הזו בין שני המבחנים חשובה, ולא כדאי סתם לבחור את המספר הנוח. מכיוון שמבחן בארטלט כבר אמר לנו שההנחה של ה-ANOVA עצמו מופרת, התשובה האמינה יותר כאן היא זו של קרוסקל-וואליס: על סכום הטיפ הגולמי, ארבעת הימים כנראה כן שונים במידה מסוימת, רק לא בצורה שמבחן רגיש-לשונות עם הנחה שלא מתקיימת יכול לזהות באמינות. זו סיבה אמיתית וניתנת לבדיקה להעדיף את המסקנה של מבחן אחד על פני אחר, על אותם נתונים בדיוק, לא רק הסתייגות תיאורטית.

הצד הבייסאני של אותה שאלה פשוט יותר: מכיוון שהמודל ההיררכי כבר נותן לנו פוסטריורי מלא לכל אחד מפרמטרי הימים, אפשר לקרוא שאלת הבדל בין קבוצות ישירות מהם, בלי לבחור בין מבחנים מתחרים ובלי לדאוג לגבי איזה מהם עומד בהנחותיו. בדקתי את ההפרש בין השיפוע של חמישי לשיפוע של ראשון. לפני שממשיכים, הערה מתודולוגית שחשוב לומר בקול: בחרתי דווקא בזוג הזה כי הוא הראה את ההבדל הגדול ביותר בטבלת הסיכום, לא כי קבעתי מראש שאבדוק אותו. יש בזה סיכון אמיתי: אם בעצם ״בדקנו בעין״ את כל שישה הזוגות האפשריים בין ארבעת הימים ובחרנו להריץ מבחן פורמלי רק על הזוג שנראה הכי קיצוני, אז ערך מובהקות בודד על הזוג הזה מזלזל בסיכוי שפער כזה יצוץ במקרה איפשהו מתוך כל הזוגות. התיקון הקלאסי הסטנדרטי למצב כזה נקרא תיקון להשוואות מרובות, וכולל שיטות כמו תיקון בונפרוני או שיטת בנג׳מיני-הוכברג הפחות שמרנית, שמיושמות על כל הזוגות שבאמת נבדקו. שווה לקרוא לזה בשמו במפורש במקום לדווח בשקט ערך מובהקות אחד מחמיא כאילו נקבע מראש. רווח האמון של 95 אחוז על ההפרש יוצא בקושי חיובי לגמרי, כמעט נוגע באפס בקצה התחתון שלו, וההסתברות הפוסטריורית ששיפוע חמישי גדול משיפוע ראשון יוצאת קרוב לתשעים ושמונה אחוז: עדות אמיתית וסבירה, אבל לא מוחלטת, לא ברמת הביטחון שראינו כשהשווינו שני מחוזות בית ספר

שונים מאוד במחברת הקודמת. זו בדיוק הרמת ביטחון הנכונה לדווח עליה: יש כאן הבדל אמיתי בקשר בין חשבון לטיפ בין שני הימים האלה, רק לא הבדל מוחץ. תמונה מכוילת יותר ממה שתווית "מובהק / לא מובהק" בודדת הייתה נותנת.

סיכום

לסיכום הכול: ברגרסיה הבסיסית, OLS ובייסאני יצאו כמעט זהים במספרים; בתחזית, אותה צורה איכותית של רצועות; בהיררכיה, שני המודלים הקלאסי והבייסאני מכווצים קבוצות קטנות באופן דומה מאוד; ובשאלת ההבדל בין ימים, ANOVA וקרוסקל-וואליס נתנו תשובות סותרות שנפתרו בעזרת בדיקת הנחה, בעוד הפוסטריורי הבייסאני נתן ישר תשובה מכוילת בלי הצורך לבחור מבחן מראש. ורק במודל הבייסאני, ולא בקלאסי המקביל לו, הייתה בכלל בעיית דגימה כמו הדיברגנסים, וגם הפתרון שלה.