קואינטגרציה
(Cointegration)

עידו גרינברג
מרץ 2015

תוכן עניינים

- מהו תהליך סטציונרי (בגדול: לא משתנה בזמן)
- בעיות בניתוח סטטיסטי של תהליכים לא סטציונריים (בגדול: עד שסיימנו לאסוף נתונים, הנתונים הראשונים כבר רלוונטיים רק לעבר; וגרוע מכך)
- קואינטגרציה (בגדול: רדוקציה לבעיה סטציונרית)
- סיבתיות (בגדול: מה הביצה ומה התרנגולת?)

מוטיבציה – בעיות לדוגמה

- הטית המצב הקיים ומייצבים אוטומטים במשק הישראלי / ד״ר ירון זליכה (החשב הכללי במשרד האוצר) ואבישר כהן, 2007
  - מאמר המסביר כיצד יש להתאים את מדיניות המס למדיניות המאקרו-כלכלית של הממשלה, כדי לעודד או לרסן פעילות במשק בצורה אפקטיבית

  *לצורך בדיקת הקשר בין משתנים שאינם **סטציונרים** יש להריץ מודל הבוחן את הקשר ארוך הטווח בין המשתנים (**קואינטגרציה**) ולאחר מכן מודל לתיקון הטעויות בטווח קצר (ECM). שיטה זו ידועה כ**שיטת אנג׳ל גריינג׳ר** שהציעו תהליך אמידה בשני שלבים. שלב ראשון אמידה של משוואת קואינטגרציה.*

- Cointegration of event-related potential (ERP) signals in experiments with different electromagnetic field (EMF) conditions / Health, 2010
  - מחקר של תגובות המוח לגירויים (מה שמודדים בבדיקת EEG סטנדרטית)

  Due to their non-stationarity, ERP signals are difficult to study. The concept of cointegration might overcome this problem and allow for the study of the co-variability between whole ERP signals…

סדרת זמן .VS נתוני חתך

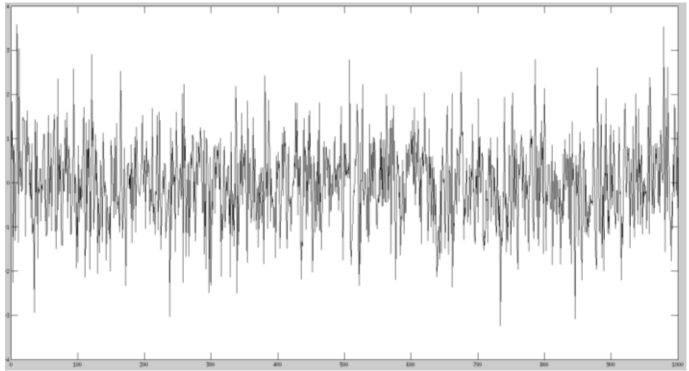
- **נתוני חתך רוחב** (*cross sectional data*) = אוסף נתונים המתארים פרטים שונים בנקודת זמן מסוימת (או ללא חשיבות לזמן)
  - דוגמאות:
    - התפלגות הגובה באכלוסייה
    - הכנסות של חברות שונות במשק
    - אחוז ההצבעה בערים שונות
- **סדרת זמן** (*time series*) = סדרת נתונים המתארת תהליך מתפתח בזמן
  - דוגמאות:
    - גובה של אדם בגילאים שונים
    - המיקום של כדור הנזרק אופקית
    - ערך מניה כתלות בזמן
  - מתמטית: תהליך סטוכסטי = סדרה של משתנים מקריים X_i (האינדקסים i מייצגים נקודות זמן שונות)
- בהרצאה זו – סדרות זמן בלבד

סטציונריות

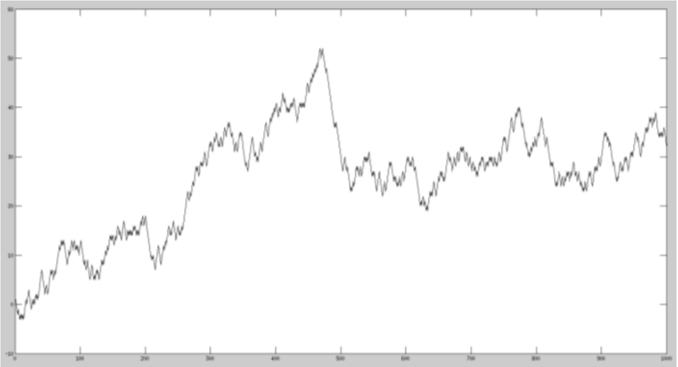
- **תהליך סטציונרי** = תהליך סטוכסטי עם התפלגות קבועה בזמן
- פורמלית:

$$F_X(x_{t_1+\tau}, \ldots, x_{t_k+\tau}) = F_X(x_{t_1}, \ldots, x_{t_k})$$

  - (קיימות גם הגדרות חלשות יותר)
- דוגמה: רעש לבן
  - (סדרת מ"מ עם התפלגות אחידה על פני קטע וקבועה בזמן)

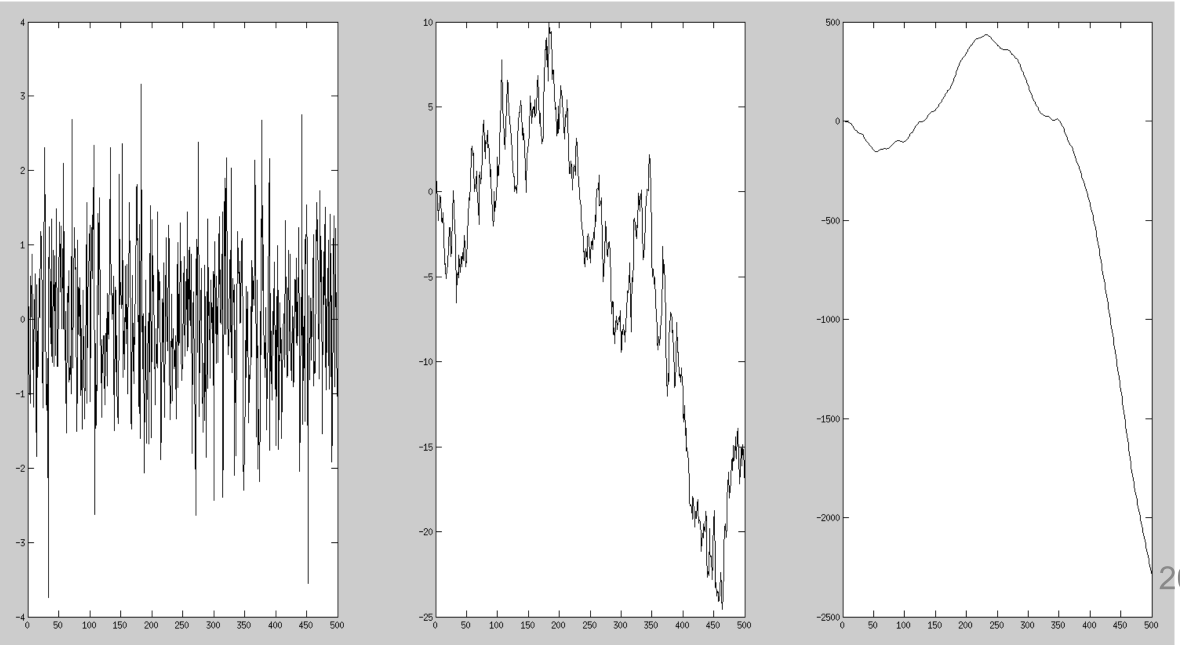


- תהליכים לא סטציונריים:
  - הגובה הממוצע באוכלוסייה
  - הילוך שיכור $X_{t+1}=X_t+u_t$
    - למרות שהשינוי בכל צעד $\Delta x_i = u_t$ כן סטציונרי



סדרות אינטגרטיביות (integrated series)

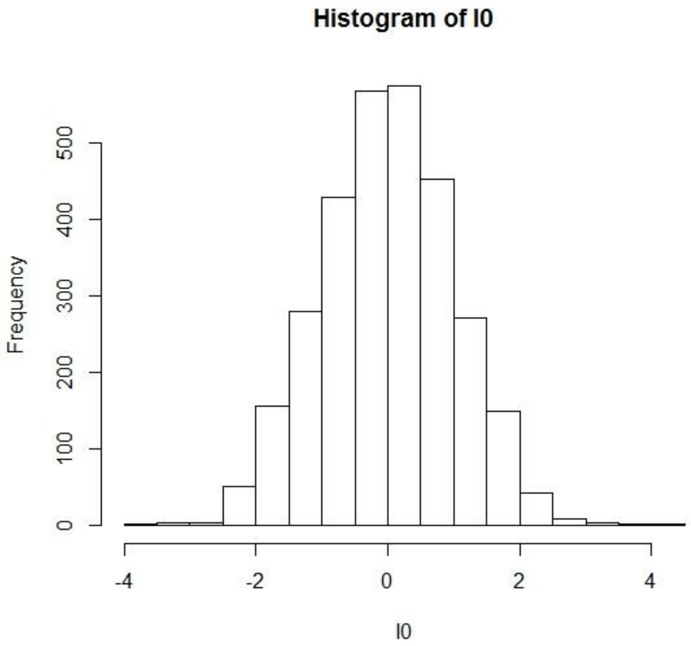
- **סדרה אינטגרטיבית מסדר n** = ההפרשים מסדר n הינם סטציונריים
- סימון: I(n)
- דוגמה – נהג משוגע שלוחץ על הבלם בכל שניה בעוצמה רנדומלית:
  - התאוצה סטציונרית
  - המהירות אינטגרטיבית מסדר 1
  - המיקום אינטגרטיבי מסדר 2
- דוגמה: הילוך שיכור אינטגרטיבי מסדר 1

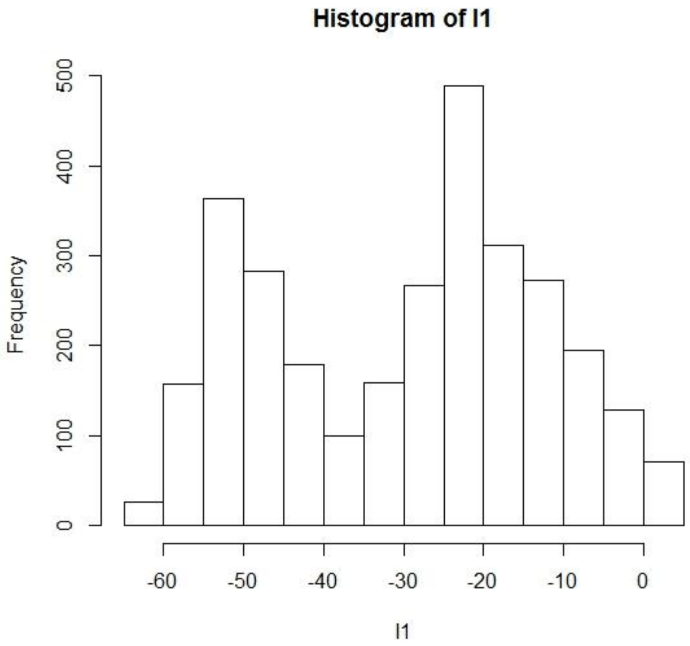


חשיבות הסטציונריות

- התכנסות דגימות לממוצע
- הערכת ההתפלגות מתוך דגימות מזמנים שונים

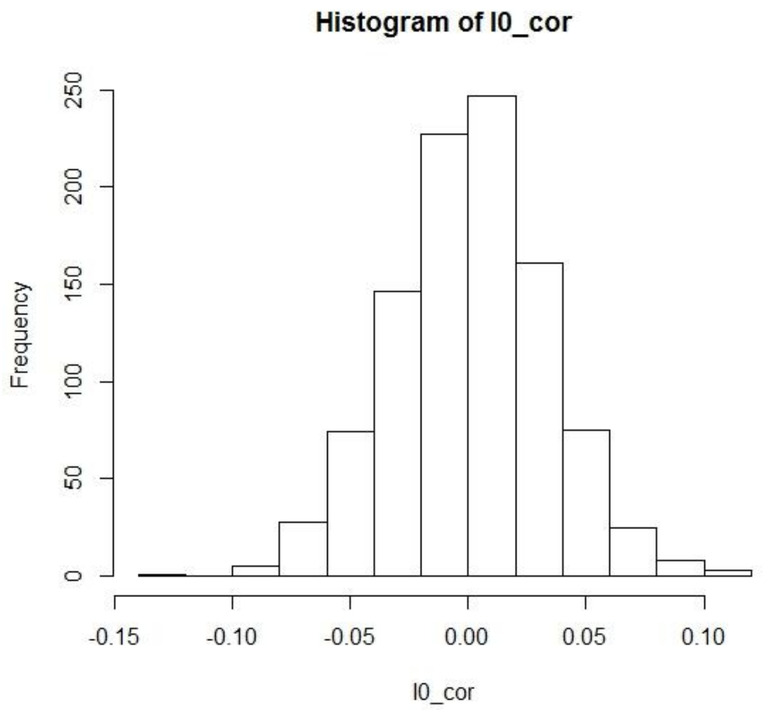
“The practical problem facing econometricians is… to find any relationships that survive long enough to be useful”

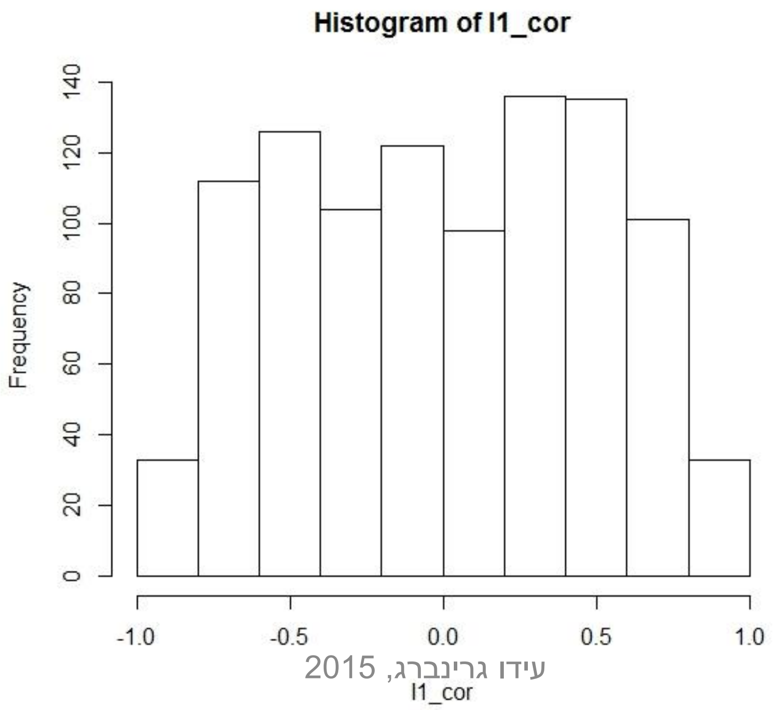


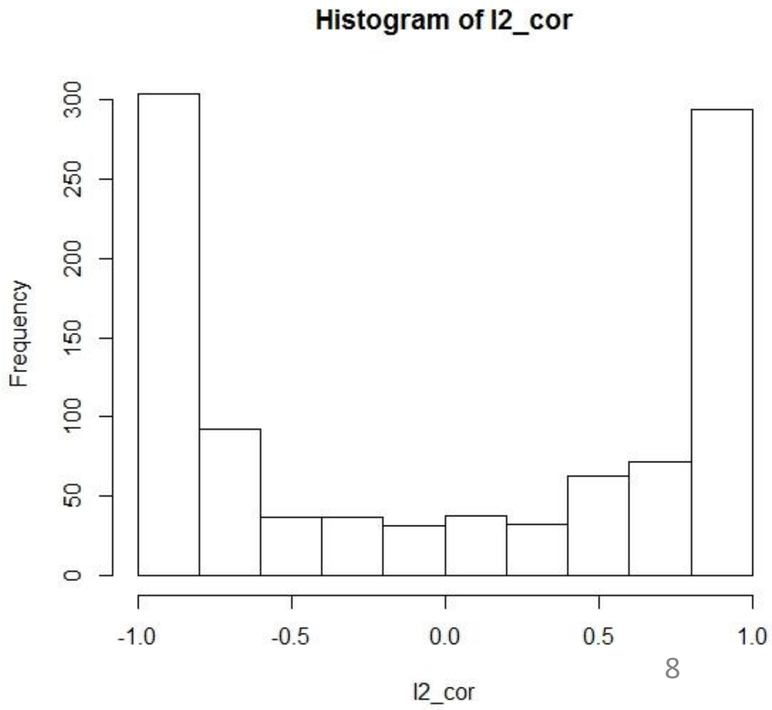


הניסוי של יול (Yule, 1925)

- אם Y,X משתנים מקריים ב"ת (iid), הקורלציה ביניהם צפויה להיות $r \approx 0$.
- נגריל 1000 זוגות של סדרות כאלה בנות 1000 איברים:
- עכשיו נגריל זוגות של תהליכים I(1),I(2) (עדיין ב"ת):

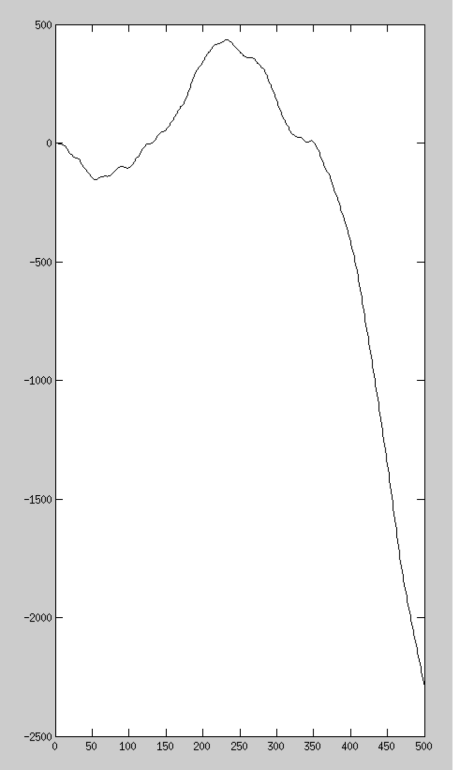


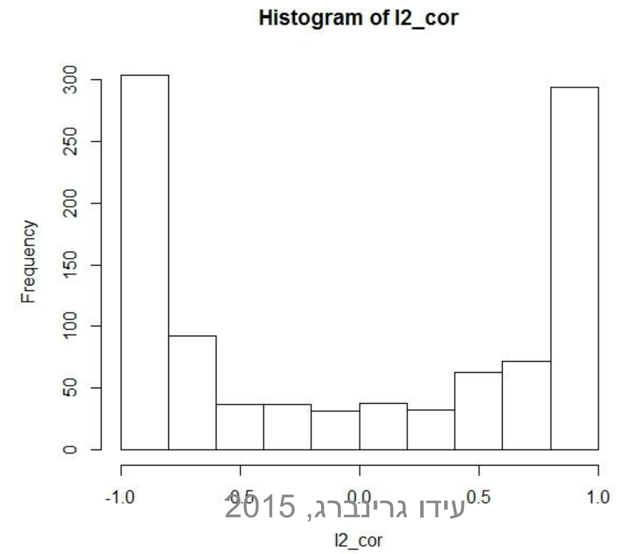




הניסוי של יול – רגרסיה מזויפת

- **רגרסיה מזויפת** (*spurious regression*) = רגרסיה שרירותית בין משתנים לא סטציונריים, שאינם בעלי קשר סיבתי
- אינטואיציה לדוגמה: משתנים I(2) נוטים "לתפוס מגמה" – לעלות או לרדת על פני מקטעים שלמים (כי ה"נגזרת" היא I(1) ולכן "רציפה")
- לכן מתקבלת קורלציה (חיובית או שלילית) גבוהה בין שני משתנים כאלה





רגרסיה מזויפת – דוגמה

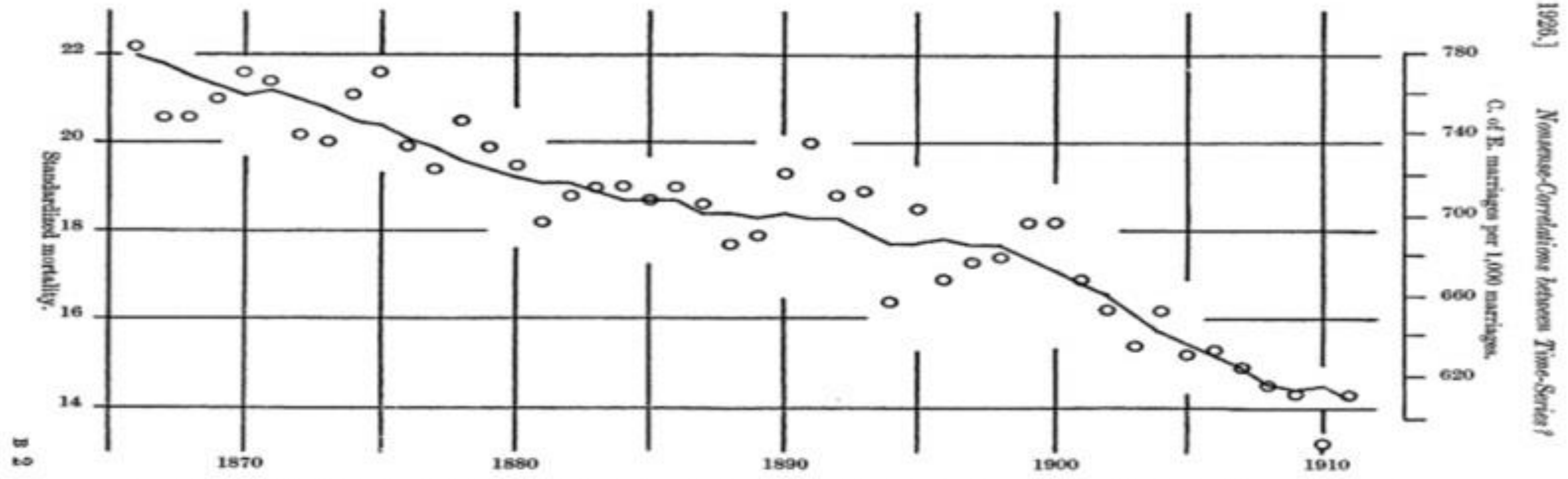


FIG. 1.—Correlation between standardized mortality per 1,000 persons in England and Wales (circles), and the proportion of Church of England marriages per 1,000 of all marriages (line), 1866–1911. $r = +0{\cdot}9512$.

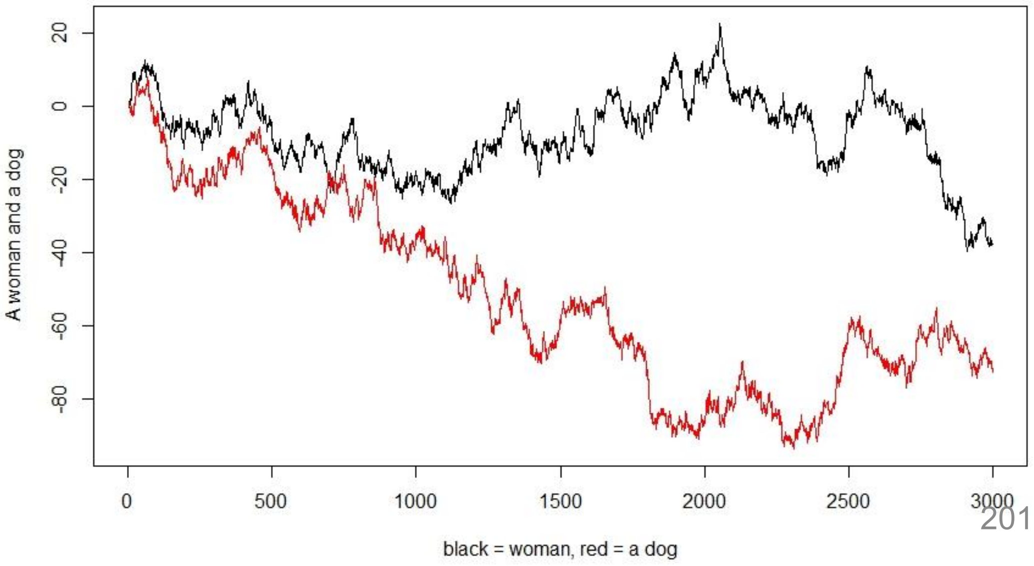
- קורלציה ~ 0.95
- קו = שיעור החתונות ב-Church of England מכלל החתונות באנגליה באותה שנה
- נקודות = שיעור המוות באוכלוסייה באותה שנה

בעיות ברגרסיה ליניארית של תהליכים לא סטציונריים

- Yule – רגרסיה מזויפת – זיהוי שגוי של רגרסיה חסרת משמעות סיבתית
- פיתרון אפשרי: לבחון קורלציה בין ההפרשים מסדר n (שהינם סטציונריים)
  - בעיה: לא ניתן לזהות מנגנונים לתיקון שגיאות *(error correction mechanism)*
  - אסביר לכם בדרך המשל

משל השיכור והכלב

- הילוך שיכור: $X_{t+1}=X_t+u_t$ (u_t רעש לבן)
  - תוחלת 0: המקום האחרון בו השיכור נראה, הוא המקום הסביר ביותר למצוא אותו
  - שונות גדלה בזמן: ככל שעובר זמן רב יותר, כך סביר יותר שהשיכור יתרחק
- נניח שיכור וכלב המבצעים כל אחד תנועה שרירותית וחסרת מטרה של הילוך שיכור, ללא קשר ביניהם:

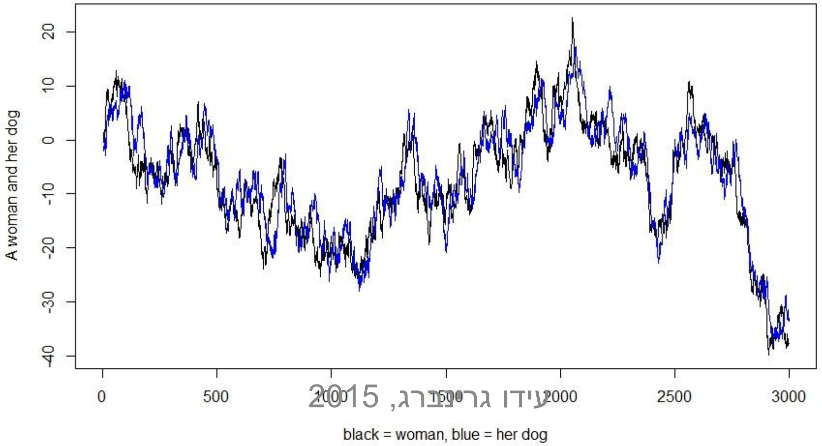


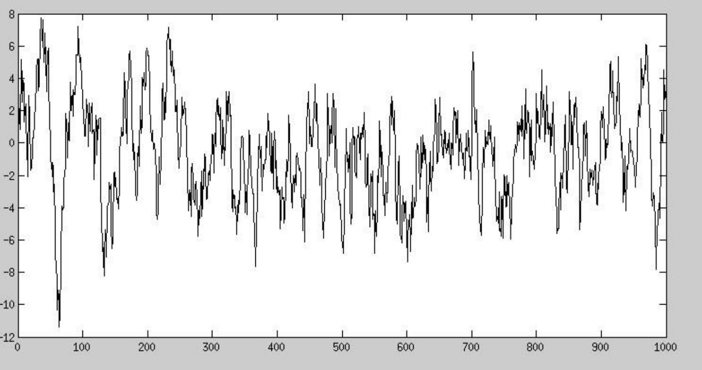
משל השיכור והכלב

- כעת נניח שהכלב והשיכור – נוסף על השיטוט המקרי שלהם – שואפים לחזור זה אל זה בכל צעד זמן:

$$X_t - X_{t-1} = u_t + c(Y_{t-1} - X_{t-1})$$
$$Y_t - Y_{t-1} = w_t + d(X_{t-1} - Y_{t-1})$$

- זהו מנגנון לתיקון שגיאות *(error correction mechanism)*
- כל אחד עדיין מבצע תנועה לא סטציונרית, אם כי לא הילוך שיכור
- כך זה נראה:





משל השיכור והכלב

- ננסה לבנות מודל ליניארי סטטיסטי לקשר בין השיכור והכלב:
  - רגרסיה ליניארית למסלולים אינה אמינה כי שניהם לא סטציונריים
  - צעדי הזמן $\Delta X_t, \Delta Y_t$ כן נראים סטציונריים
    - אבל חישוב של הקורלציה ביניהם נותן 0.03 בלבד, למרות שיש קשר ניכר בין המסלולים
    - 🡐 הרדוקציה לבעיה הסטציונרית של הפרשי הזמן איבדה את המידע על מנגנון תיקון השגיאות
  - ננסה לשאול משהו אחר – מה המרחק בין השיכור והכלב $X_t - Y_t$?
    - המרחק נראה סטציונרי 🡐 אפשר לחקור את ההתפלגות שלו (בתנאי שנאושש את הסטציונריות)
    - ניתן להבין את הקשר בין המשתנים, גם אם לא ניתן לחזות את ההתפתחות של כל אחד בנפרד
    - במקרה זה, המרחק בין השיכור והכלב מכונה **יחס קואינטגרציה** (*cointegration relationship*)

קואינטגרציה

- באופן כללי:
  - אוסף תהליכים סטוכסטיים אינטגרטיביים מסדר n הינם **קו-אינטגרטיביים** (*cointegrated*) אםם קיים צירוף ליניארי שלהם שהינו סטציונרי
  - במקרה זה, הצירוף הליניארי ייקרא **יחס קואינטגרציה** (*cointegration relationship*)
- יחס קואינטגרציה הוא צ"ל "יוצא דופן" (ולעיתים יחיד) – "סתם סכום" של המשתנים יישאר לא סטציונרי!
- Engle & Granger (1987):
  - תחת מספר תנאים סטנדרטיים, תהליכים סטוכסטיים הינם קואינטגרטיביים אםם קיים בהם מנגנון לתיקון שגיאות

סיכום ביניים

- תהליך סטציונרי – עם התפלגות קבועה בזמן
- חשיבות הסטציונריות:
  - אוסף דגימות מזמנים שונים מתכנס להתפלגות ה"אמיתית"
  - רוב המודלים הסטטיסטיים (כגון רגרסיה ליניארית) מסתמכים על סטציונריות
- בהיעדר סטציונריות:
  - רגרסיה ליניארית נותנת תוצאות מטעות וחסרות משמעות (Yule)
  - רדוקציה להפרשי זמן סטציונריים – הפרשים בין צעדי זמן – מאבדת מידע חיוני (משל השיכור והכלב)
  - קואינטגרציה – רדוקציה להפרשים סטציונריים בין המשתנים השונים – מאפשרת למדל את הקשר ביניהם

קואינטגרציה – אז מה עכשיו?

- איך מזהים קואינטגרציה?
  - **מבחן Engle-Granger – האם קיים צ"ל סטציונרי?**
    - **מבחן Dickey-Fuller – האם משתנה הוא סטציונרי?**
- איך מוצאים יחס קואינטגרציה?
  - רגרסיה ליניארית
- מה עושים עם זה?
  - חוקרים את ההתפלגות של יחס הקואינטגרציה
  - משתמשים בשיווי המשקל של המערכת ובנטייה לחזור אליו
  - יכולים להשתמש במודל ליניארי עבור המשתנים ללא חשש מרגרסיה מזויפת
    - $\alpha x + \beta y$ סטציונרי ← רגרסיה ליניארית תיתן $y \approx -\frac{\alpha}{\beta} x$, אחרת השארים יתבדרו

שורש יחידה (*Unit Root*)

- לתהליך $x_t = \rho x_{t-1} + u_t$ (u_t are iid) יש **שורש יחידה** אםם $\rho = 1$
  - מקור השם בתהליך המוכלל $x_t = a_1 x_{t-1} + \cdots + a_d x_{t-d} + u_t$, שם יש שורש יחידה אםם m=1 הוא שורש של הפולינום האופייני $m^d + a_1 m^{p-1} + \cdots + a_d$
  - כמובן שאם יש שורש יחידה אז התהליך הוא I(1)
  - נשים לב שלכל $\rho < 1$, מתקיים $x_n = \sum_{t=1}^{n} \rho^{t-1} u_t \approx \sum_{t=n-d}^{n} \rho^{t-1} u_t$, ולכן x_n סטציונרי בקירוב
- <u>בדיקת סטציונריות</u>: בהינתן סדרה $x_1, \ldots, x_n$, האם יש שורש יחידה?
  - האם $\rho = 1$?
- **נציג כ-$\Delta x_t = \theta x_{t-1} + u_t$, ונשאל: האם $\theta = 0$?**

מבחן Dickey-Fuller לקיום שורש יחידה

$$\Delta x_t = \theta x_{t-1} + u_t$$

- האם $\theta = 0$?
- נתאים קו מהצורה $\Delta x_t = \hat{\theta} x_{t-1}$ ($\hat{\theta}$ זו ההתאמה האופטימלית לנתונים לפי Least Squares)
- נניח שהתקבל $\hat{\theta} = -0.01$. האם בפועל $\theta = 0$ או $\theta < 0$?
- ננרמל $DF = \frac{\hat{\theta}}{SE(\hat{\theta})}$, ונשאל: בהנחה ש-$\theta = 0$, מה הסיכוי לקבל את $\hat{\theta}$ שיצא או רחוק יותר מ-0?
- זהו מבחן סטטיסטי – **בדיקת השערות** – בודקים בשלילה האם ההשערה $\theta < 0$ נכונה
- הקישור בין ערך של $\hat{\theta}$ לבין הסיכוי לקבלו כאשר $\theta = 0$, מתקבל מ**טבלת Dickey-Fuller**

<u>בשורה התחתונה</u>: **אנחנו יודעים לבדוק סטטיסטית האם תהליך הוא סטציונרי**

מבחן Engle-Granger לקואינטגרציה

- שני תהליכים $\{x_t\},\{y_t\}$ (או יותר) הם קואינטגרטיביים אםם קיים צ"ל סטציונרי
- מבחן Engle-Granger:
  - מחשבים רגרסיה ליניארית: $y_t \approx m \cdot x_t$
  - מחשבים שארים (השגיאות של המודל הליניארי): $r_t = y_t - m \cdot x_t$
  - מבחן Dickey-Fuller: האם השארים הם תהליך סטציונרי?
    - אם כן: מצאנו צ"ל סטציונרי ➔ **יש קואינטגרציה!**
    - אם לא: לא מצאנו צ"ל סטציונרי. האם בכל זאת יש קואינטגרציה?
      - נניח בשלילה שכן, ונקבל שהרגרסיה הליניארית מייצגת (עד כדי שגיאה סטטיסטית) את יחס הקואינטגרציה, ולכן השארים סטציונרים ➔ סתירה ➔ **אין קואינטגרציה!**

סיבתיות

It is difficult to make predictions, especially about the future (Niels Bohr)

- עיקרון הסיבתיות של רייכנבאך: אם קיימת קורלציה בין X ל-Y, אז:
  - X משפיע על Y, או
  - Y משפיע על X, או
  - קיים Z המשפיע על X ו-Y
- אז איך תיתכן קורלציה בין תהליכים לא סטציונריים גם ללא קשר סיבתי?
  - ההסבר: הזמן הוא התהליך הנוסף המשפיע על השניים האחרים
  - לדוגמה (Yule ידידנו): הקשר בין שיעור החתונות הקתוליות לשיעור התמותה בבריטניה – שניהם פוחתים עם הזמן

סיבתיות גריינג'ר (*Granger Causality*)

- האם X משפיע על Y?
  - תנאי הכרחי – **סיבתיות גריינג'ר**: ניבוי של Y ע"י העבר של Y, <u>טוב פחות</u> מניבוי של Y ע"י העבר של X ושל Y ביחד
- מבחן גריינג'ר:
  - רגרסיה של Y_t לפי ערכי עבר של Y וערכי עבר של X:

$$Y_t = \sum_{j=1}^{d} a_j X_{t-j} + \sum_{j=1}^{d} b_j Y_{t-j} + r_t$$

  - האם X תרם מידע משמעותי? האם $a_j \not\equiv 0$?
    - שוב בדיקת השערות: נניח בשלילה ש-$a_j \equiv 0$. מה הסיכוי לקבל את המקדמים האמפיריים a_j שחושבו?
    - הסיכוי ניתן ע"י משתנה סטטיסטי בשם Wald Statistic

סיבתיות גריינג'ר וקואינטגרציה

- מגבלות של מבחן סיבתיות גריינג'ר:
  - נבדק רק קשר ליניארי
  - מניח ש-X,Y תהליכים סטציונריים
    - או קואינטגרטיביים
- <u>משפט</u>: אם X,Y קואינטגרטיביים, אז לפחות אחד מהם משפיע על השני (במובן של סיבתיות גריינג'ר)
- דוגמה: נגריל 2 סדרות:

$$X_t - X_{t-1} = u_t$$
$$Y_t - Y_{t-1} = w_t + 0.05 \cdot (X_{t-1} - Y_{t-1})$$

ממבחן גריינג'ר התקבל:

- בהנחה ש-X אינו משפיע על Y, הסיכוי לקבל את הסדרות הנ"ל: 2e-14
- בהנחה ש-Y אינו משפיע על X, הסיכוי לקבל את הסדרות הנ"ל: 0.27

מקורות

- מצגת משותפת של ליאת שנהב ושלי מ-2014

- Michael P. Murry: A Drunk and Her Dog

http://www.uta.edu/faculty/crowder/papers/drunk%20and%20dog.pdf

- David F. Hendry, Katarina Juselius: Explaining Cointegration Analysis: Part I

http://www.econ.ku.dk/okokj/papers/dfhkjfnl.pdf

- Soren Johansen: Correlation, regression, and cointegration of nonstationary economic time series

http://www.math.ku.dk/~sjo/papers/LisbonPaper.pdf