הלשכה המרכזית
לסטטיסטיקה

מס' 15 .NO

שיטה להחלקת פונקצית התמותה באמצעות מודל רגרסיה בקטעים:
יישום על נתונים ישראלים

A Method for Smoothing Mortality Functions using a segmented regression model:
an application to Israeli data

דר' אלברט וקסלר, גב' נטלי פלאקס-מנוב ומר ארי פלטיאל

Albert Vexler, Natalie Flaks-Manov, Ari Paltiel

חשון, תשס"ו, דצמבר, December 2005

הלשכה המרכזית לסטטיסטיקה – גף בריאות

Central Bureau of Statistics - Health Division

רחוב כנפי נשרים 66 פינת רחוב בקי, גבעת שאול, ת"ד 13015, ירושלים 95464 טלפון: 02-6592666, פקס' 02-6521340

דואר אלקטרוני: info@cbs.gov.il כתובת האתר: www.cbs.gov.il

מודל לחישוב לוח תמותה שלם בישראל

דר׳ אלברט וקסלר, גב׳ נטלי פלאקס-מנוב ומר ארי פלטיאל

תקציר

נייר זה מציג שיטה חדשה להחלקת פונקצית התמותה כדי לאמוד לוח תמותה שלם. בהסתמך על נתונים ישראלים, נעשתה השוואה בין תוצאות המודל לבין החלקה באמצעות ״חוק התמותה״ שהוצע על ידי Heligman and Pollard (1980), אשר יושם בחבילת התוכנה של האו״ם – Mortpak. השיטה החדשה מבוססת על חלוקת פונקצית התמותה לשני חלקים ע״י רגרסיה דו-שלבית ואמידת נקודת השבר של המודל. בין היתרונות של השיטה החדשה הם פשטות, נוחות היישום ואספקת פרמטרים סטטיסטיים (שונות, רווח סמך) לנקודות על פונקצית התמותה המוחלקת. השיטה מבטיחה שההבדלים בין תוחלת החיים המחושבים מהנתונים האמפיריים הגולמיים ואלה המחושבים מהפונקציה המוחלקת לא יהיו מובהקים סטטיסטית.

Abstract

The paper presents a new method for smoothing a mortality function in order to estimate a complete life table. Using Israeli data, results of the model are compared to graduation using estimation of the "mortality law" proposed by Heligman and Pollard (1980), as implemented in the United Nations software package *Mortpak*. The new model divides the mortality function into two sections, based on a two-phase regression model and an estimation of the change-point of the model. Among the advantages of the proposed method are its simplicity, the ease with which it can be implemented, and the provision of statistical parameters (variance, confidence intervals) for the smoothed points on the mortality function. The method ensures that the differences between life expectancy values calculated from the raw empirical data and those based on the smoothed function are not statistically significant.

מילות מפתח: לוח תמותה, תוחלת חיים, החלקת פונקצית תמותה, רגרסיה בקטעים.

1. מבוא

שיטות לחישוב לוח תמותה קיימות כבר מאמצע המאה ה-18. מודלים שונים פותחו ע״י אקטוארים, דמוגרפים ומומחים אחרים, על מנת לאמוד תמותה כפונקציה של גיל. המודלים המוכרים בספרות בנושא לוחות תמותה הם של Thiele מהשנים 1871-1872 (ראה Hartmann (1987)) , של Wittstein משנת 1883 (ראה Hartmann (1987)), Heligman and Pollard (1980), וכן פיתוח של Kostaki (1992), אשר הציעה שיטה לאמוד את הפרמטרים של המודל H-P בשיטת ריבועים פחותים ולשפר את המודל על סמך הוספת פרמטר נוסף.

בנייר זה מוצג מודל חדש לחישוב לוח תמותה שלם (נתונים לגיל בודד). השיטה מבוססת על חלוקת פונקצית תמותה לשני חלקים ע״י רגרסיה דו-שלבית ושיטת אמידה של נקודת השבר של המודל. השיטה מאפשרת בדיקת מובהקות הפרמטרים ובניית רווח סמך עבורם.

1.1 סקירת ספרות – שיטות קיימות לחישוב לוח תמותה שלם

לפי Hartmann (1987), הראשון שפרסם מודל ללוח תמותה לכל הגילים הוא Thiele בשנת 1872 (מודלים קודמים התאימו רק לגילים המבוגרים). השערתו העיקרית הייתה, שפונקצית תמותה מתחלקת לשלושה חלקים: גיל הילדות, אמצע החיים וגיל מבוגר. השערה זו נשארה תקפה עד היום ורוב המודלים של פונקצית תמותה בנויים משלושה חלקים.

Thiele הציע מודל מסוג:

$$\mu(x) = \mu_1(x) + \mu_2(x) + \mu_3(x)$$

כאשר $\mu(x)$ היא פונקצית הסיכון (Hazard function) - ההסתברות הרגעית למות, או במילים אחרות, ההסתברות שאדם ימות בטווח זמן רגעי, בהינתן שהוא שרד עד לתחילת אותו טווח זמן. צורת פונקצית ההסתברות משתנה בשלושה טווחי גיל:

Childhood: $\mu_1(x) = a_1 exp(-b_1 x)$

Middle life: $\mu_2(x) = a_2 exp\left(-0.5 b_2 (x-c)^2\right)$

Adult: $\mu_3(x) = a_3 exp(b_3 x)$,

כאשר שבעת הפרמטרים $a_i, b_i\ (i = 1,2,3)\ c$ חיוביים.

המודל של Thiele ניבא בצורה סבירה מאוד את התמותה בגילים המבוגרים, אך לא נתן ניבוי טוב עבור התמותה בילדות ובגילים הצעירים (תקופת התאונות וסיבות חיצוניות אחרות).

לפי Hartmann (1987), כבר ב-Elston 1923 סקר את כל חוקי התמותה הקיימים ומצא, שהמאפיין המרכזי של מודלים אלו הוא ניבוי טוב של פונקצית התמותה בגילים המבוגרים וכישלון בניבוי פונקצית התמותה בגיל הילדות ובמיוחד בתקופת התאונות.

מאוחר יותר פותח המודל של Heligman-Pollard ((e.g. Heligman and Pollard (1980), אשר הציעו מודל המורכב משמונה פרמטרים:

$$q_x / p_x = A^{(x+B)^C} + Dexp\left(- E\left(ln\frac{x}{F} \right)^2 \right) + GH^x,\ p_x = 1 - q_x$$

A, B, C, D, E, F, G, H – הם פרמטרים חיוביים, כאשר כל איבר בנוסחה מבטא את התנהגות הפונקציה בטווח גיל אחר. q_x הוא ההסתברות למות בגיל x, ו- p_x הוא המשלים, ההסתברות לחיות.

המודל של H-P נותן התאמה טובה לנתונים אמפיריים. מודל זה מתואר בספרות כשימושי ביותר לצורך חיזוי תמותה בתחזיות אוכלוסייה לסימולציות דמוגרפיות, בהן דרושים הנתונים של תמותה בכל גיל. תוכנה לחישוב המודל אף הוכנסה לחבילת התוכנות הסטנדרטית של האו״ם –Mortpak (UN 1988). בישראל השתמשו בתוכנה זו עד לשנת 2002, אך משנת 2001 ניתן היה להבחין, כי השיטה של H-P גם היא נכשלת בניבוי פונקצית התמותה של האוכלוסייה בישראל בגילי הילדות וכן בתקופת התאונות. כמו כן, קיימת גם בעיה בהתאמת המודל להסתברויות למות בגילים המאוד מבוגרים (+80). בפרק הבא נפרט יותר על הקושי להשתמש בשיטה של H-P בישראל.

המודלים האחרונים של פונקצית תמותה המופיעים בספרות הם של Kostaki (1992). היא מתבססת על המודל של H-P ומנסה לשפר אותו ע״י הוספת פרמטר תשיעי למודל.

$$\frac{q_x}{p_x} = \begin{pmatrix} A^{(x+B)^C} + Dexp\left(- E_1\left(log\frac{x}{F} \right)^2 \right) + GH^x, x \le F \\ A^{(x+B)^C} + Dexp\left(- E_2\left(log\frac{x}{F} \right)^2 \right) + GH^x, x > F \end{pmatrix}$$

הוספת פרמטר תשיעי למודל ממזער טוב יותר את סכום הריבועים של ההפרשים בין המודל ונתונים אמפיריים וכן תורם לניבוי סביר יותר של ההסתברות למות בגיל התאונות ובגילים הצעירים. אך, הוספת פרמטר למודל מצריכה אמידה של הפרמטרים בעזרת נוסחה לא ליניארית, המבוססת על נתונים אמפיריים. על מנת לאמוד את המובהקות של הפרמטר צריכים לבנות רווח סמך עבורו. הספרות המתארת את אמידת הפרמטרים בנוסחאות של Kostaki לא מראה רמת מובהקות של הפרמטר התשיעי. מכאן, שימוש בנוסחאות Kostaki ללא התייחסות למובהקות האומדים יכול להוביל למסכנות מוטעות.

1.2 סיבות למעבר לשיטת חישוב חדשה בישראל

שיעורי התמותה בישראל, בדומה למדינות אחרות בעולם, חשופים לסטיות אקראיות (טעויות סטטיסטיות) ולסוגים שונים של טעויות שאינן סטטיסטיות, כגון אלה הנובעות מדיווח שגוי של שנת לידה או של גיל בעת הפטירה. שני סוגי הטעויות גורמים לכך, ששיעורי התמותה המחושבים שונים משיעורי התמותה ה״אמיתיים״ שהיינו מחשבים לו ניתן היה להתגבר על הטעות הסטטיסטית והטעות בדיווח. הטעויות הסטטיסטיות הן משמעותיות יותר ככל שמדובר בקבוצת אוכלוסייה קטנה יותר, בקבוצת גיל

בודדת או בתקופה קצרה יותר, דבר שעלול להוביל להתנהגות לא סדירה של הנתונים האמפיריים. כדי להתגבר על בעיה זו נהוג להשתמש בשיטת "החלקה" מסוג כלשהו.

לוח תמותה "מקוצר", המבוסס על שיעורי תמותה של קבוצות גיל רחבות (ולא גיל בודד) חשוף פחות לסטיות אקראיות. הבעיות חמורות יותר בחישוב לוח תמותה "שלם", המבוסס על גיל בודד.

לוחות התמותה השלמים בישראל לשנים 1986 ואילך חושבו בעזרת תכנת MORTPAK, שסופקה ע"י האו"ם. תוכנה זו מאפשרת הכנת לוחות תמותה שלמים על ידי אמידת מודל מסוג Heligman-Pollard (H-P) ((e.g. Heligman and Pollard (1980)), בשיטת מזעור ריבועים פחותים. בשנים האחרונות התברר שתוכנה זו אינה מספקת תוצאות סבירות לנתונים הישראליים, כלומר התאמת המודל לנתונים אמפיריים אינה מובהקת. נמצא כי נוסחת H-P מעלה את תוחלת החיים בלידה בכל קבוצות האוכלוסייה (לפחות ב-0.2 שנים) לעומת לוח התמותה המקוצר. בגילים המבוגרים ההבדלים בין שיטת H-P לעומת לוח תמותה מקוצר גדולים עוד יותר. בקרב זכרים באוכלוסייה הערבית למשל, השיטה של H-P מורידה את תוחלת החיים בגיל 80 כמעט בשנתיים. הקו של המודל חורג מגבולות רווח הסמך של נתוני ההסתברות למות (q_x) האמפיריים. כמו כן, הפרמטרים של נוסחת H-P ניתנים לאמידה, אך לא ניתן לחשב את המדדים הסטטיסטיים (סטיית תקן ומובהקות) של אומדני הפרמטרים, לכן לא ידועה לנו רמת המובהקות של המודל. לבסוף, מודל זה מבצע החלקה שלא מבטאת את הייחודיות של הנתונים הישראליים. בגילים מסוימים ההחלקה מקטינה מאוד את ההסתברויות למות (כגון בגיל הצבא) ובגילים אחרים היא מגדילה אותם (בעיקר בגילים המבוגרים).

לפיכך פותחה שיטה חדשה שהיא קירוב לפונקצית q_x ע"י פולינום דו-שלבי. המודל מבוסס על שימוש בשיטת נראות מקסימלית מקומית (Local Maximum Likelihood)

((e.g. Fan, Farmen and Gijbels (1998)) ושימוש בשיטת אומדנים של נקודת שבר (changepoint) (e.g. Koul, Lianfen and Surgailis (2003)).

לשיטה זו ארבעה יתרונות:

- ההבדלים בין תוחלת החיים לפני החלקת הנתונים האמפיריים של הסתברויות למות בגיל בודד ולאחר ההחלקה אינם מובהקים.
- ניתן לחשב מדדים סטטיסטיים של המודל, כמו שונות, רווח סמך ומובהקות.
- המודל משיג קירוב טוב יחסית לנתונים האמפיריים ע"י החלקת ה-q_x (הסתברות למות) ומתחשב בייחודיות הנתונים הישראליים.
- השיטה קלה ונוחה לשימוש.

בשיטה החדשה חישוב תוחלת חיים מתבצע בארבעה שלבים:

שלב א': חישוב ערכי q_x על סמך שיעורי התמותה $\left(m_x\right)$ לגיל בודד לכל קבוצת אוכלוסייה ולכל מין, בממוצע לתקופה של חמש שנים (1996-2000).

שלב ב׳: בדיקת השערות על קיום נקודת שבר במודל. אם ההשערה לא נדחית עוברים לשלב ג׳.

שלב ג׳: החלקת q_x על פי אמידה של שני מודלים של פונקצית ה-q_x, אחד לגילים הצעירים (עד נקודת שבר) ושני לגילים מבוגרים יותר (אחרי נקודת השבר).

שלב ד׳: חישוב כל הפרמטרים של לוח תמותה בהתבסס על אומדני ה-q_x שהתקבלו מהמודלים.

2. שיטות חישוב

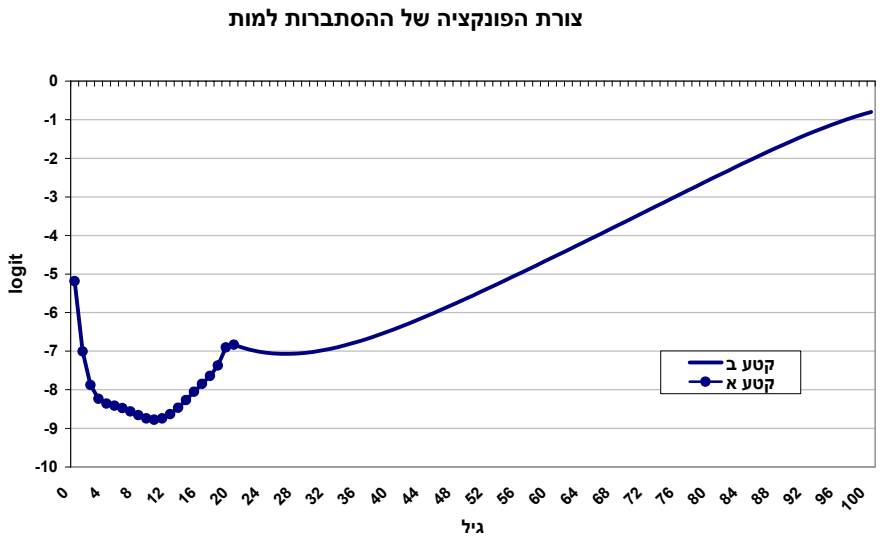
ידוע כי יחס הסיכויים למות בין גיל x ל x+1 מוגדר ע״י הנוסחה:

$$f(x) = \frac{q_x}{(1-q_x)} \quad , \qquad (1)$$

(e.g. Heligman & Pollard (1980)).

כמו כן, ידוע כי ההסתברות למות תלויה בגיל ומתנהגת באופן שונה בטווחי גיל שונים. כך למשל, פונקצית ההסתברות למות בגילים צעירים שונה מפונקציה תואמת בגילים מבוגרים. לדוגמא, מצורף גרף של פונקציה אמפירית $f(x)$, המתאר הסתברות למות של זכרים בכלל אוכלוסיית ישראל. בתרשים 1 רואים בבירור שני חלקים של פונקצית התמותה. דוגמאות נוספות לפונקציות תמותה ניתן לראות במאמרה של Kostaki (1991).

תרשים 1:



כאשר $logit = log \frac{q_x}{1 - q_x}$

המודל של H-P פותח כדי לתפוס את ההתנהגות השונה בטווחי הגיל השונים, אבל במחיר הצורך באמידת פרמטרים רבים. Kostaki (1992), כדי לשפר את ההתאמה של המודל, הציעה אף פרמטר נוסף, תשיעי. היא הצליחה לשפר במקצת את הקירוב, אך הנתונים עליהם נבנה המודל היו שיעורי תמותה של מדינות אירופה ובבדיקה שערכנו נראה, כי המודל של Kostaki מחליק יתר על המידה את הפונקציה $f(x)$ בטווחי גילים המיוחדים לאוכלוסיית ישראל (פירוט על החלקת יתר יבוא בהמשך). לכן הוחלט לבנות מודל חדש,

אשר יהיה מתאים להתנהגות פונקצית ההסתברות למות של אוכלוסייה ישראלית. המודל שנבנה הוא פולינום דו-שלבי, המבוסס על שימוש בשיטת נראות מקסימלית מקומית (Local Maximum Likelihood) ושימוש בשיטת אומדנים של נקודת שבר (change point). השיטה מתאימה להשערה שניתן לאמוד את פונקצית התמותה משני חלקים, בטווח של הגילים הצעירים ובטווח של הגילים המבוגרים, כאשר נקודת השבר ביניהם אינה קבועה ויש צורך לגלות אותה באמצעים סטטיסטיים.

2.1 תאור המודל

ניקח $y_x = log(q_x/(1-q_x))$, נניח כי פונקציה $\log(f(x)/(1-f(x)))$ מאפשרת פירוק של הפולינום עד לדרגה p (לדוגמא קירוב של טור טיילור). כמו שמקובל בשיטה של נראות מקסימלית מקומית, נציג את המודל הנתון ע"י קירוב רגרסיה

$$y_x = \sum_{j=0}^{p} a_j x^j + \varepsilon_x , \qquad (2)$$

כאשר ε_x משתנה מקרי בלתי תלוי, המתפלג נורמלית $\varepsilon_x \sim N(0,\sigma^2)$ ו-x הוא גיל $x = 0,...100$,

אז האמידה של הפונקציה $f(x)$ ב-(1) מתאפשרת ע"י אמידה סטטיסטית של מקדמי הפולינום ב- (2).

בהסתמך על מאמרים של Heligman and Pollard ו- Kostaki ניתן להניח כי הפונקציה $f(x)$ יכולה להיות לא רציפה, כלומר הפונקציה יכולה להיות שונה בטווחי גיל שונים. על מנת להתחשב בתלות פונקציונלית שונה של $f(x)$ ב- x בטווחי גיל שונים (גיל צעיר וגיל מבוגר) נבנה מודל עם נקודת שבר (change point).

נציג את הבעיה בצורה

$$y_x = \sum_{j=0}^{p} a_{Lj} x^p + \varepsilon_x , \qquad (3)$$

כאשר L=0,1 ובדיקת השערות סטטיסטיות על קיום נקודת שבר (change point) בטווח גילים נתון,

$$H_0 : \text{for all } x : L = 0 \qquad \text{versus}$$
$$H_1 : \text{if } x < x_c \text{ then } L = 0, \text{if } x \geq x_c \text{ then } L = 1,$$
$$x_c \text{ is unkown}$$

המודל מתאר מצב בו מעל לנקודת שבר בלתי ידועה x_c הפרמטרים של המודל הם a_{0j} וכאשר $x_c < n$ הפרמטרים של המודל משתנים ל- a_{1j}.

על מנת לבדוק את ההשערות, נעשה שימוש בשיטה המוצגת במאמר של Kim ו-Siegmund (Kim and Siegmund (1989)).

אם לא דוחים H_0, אז אומדים את מקדמי הרגרסיה (3) $\alpha_{00},\alpha_{01},...,\alpha_{0p}$ בשיטת הנראות המרבית בטווח תצפיות $x = 0,...100$.

בהנחה ש- H_1 נכונה, בשלב הראשון נשתמש בשיטה של Koul ((Koul, Lianfen & Surgailis (2003)), נאמוד את נקודת השבר(change point):

$$\hat{x}_c = arg\,min_k \sum_{i=1}^{k-1}(y_i - \sum_{j=0}^{p}\hat{\alpha}_j^{(1,k-1)} i^j)^2 + \sum_{i=k}^{n}(y_i - \sum_{j=0}^{p}\hat{\alpha}_j^{(k,n)} i^j)^2$$

כאשר $\hat{\alpha}_j^{(r,m)}$ אומד נראות מרבית לפרמטר α_j המבוסס על תצפיות $(y_r,...,y_m)$.

לאחר מכן, נאמוד $(\hat{\alpha}_{00},...,\hat{\alpha}_{0p})$ בשיטת הנראות המרבית על סמך תצפיות $(y_1,...,y_{x_{c-1}})$

ו- $(\hat{\alpha}_{10},...,\hat{\alpha}_{1p})$ על סמך תצפיות $(y_{x_c},...,y_n)$.

כתוצאה נקבל אומד לאיבר דטרמיניסטי עבור פונקצית התמותה

$$\hat{q}_x = \frac{exp\left(\sum_{j=0}^{p}\hat{\alpha}_{Lj}x^p\right)}{1+exp\left(\sum_{j=0}^{p}\hat{\alpha}_{Lj}x^p\right)}, \qquad (4)$$

$$\hat{y}_x = log\left(\frac{\hat{q}_x}{1-\hat{q}_x}\right) \quad L = \begin{cases} 0, if & 1 < x < x_c - 1 \\ 1, if & x \geq x_c \end{cases}$$

השימוש בשיטה המתוארת מאפשר קבלת פרמטרים עבור מודל (3)[1].

דוגמא לחישוב $\hat{y}$, כאשר $x < 20$ בקרב סה"כ זכרים:

$$\hat{y}_{10} = -5.18 - 2.49\frac{10}{1} + 1.56\frac{10^2}{2} - 0.74\frac{10^3}{3!} + 0.24\frac{10^4}{4!} - 0.048\frac{10^5}{5!} + 0.0045\frac{10^6}{6!} \cong -9$$

ומכאן, $\hat{q}_{10} = 0.00012$

דוגמא לחישוב $\hat{y}$, כאשר $x \geq 20$ בקרב סה"כ זכרים:

$$\hat{y}_{30} = -0.35 - 0.67\frac{30}{1} + 0.048\frac{30^2}{2} - 0.0023\frac{30^3}{3!} + 0.00008\frac{30^4}{4!} \cong -6.8$$

ומכאן, $\hat{q}_{30} = 0.00111$

[1] ראה לוח "מקדמי הרגרסיה ונקודות השבר" בנספח בעמ' 24 ותרשימים מתאימים 10-17 מנתוני תמותה של הלמ"ס באוכלוסיות יהודים ואחרים, יהודים, אוכלוסייה ערבית וסך הכל בשנים 1996-2000.

3. ממצאים

3.1 יישומי המודל

על מנת לבדוק את ההתאמה בין מודל (4) לנתונים אמפיריים, בשלב הראשון נמצא רווח סמך להסתברות למות בשכבת גיל x. רווח סמך ניתן לקבל בהנחת נורמליות ע"י הנוסחה

$$q_x \pm 1.96\sqrt{\frac{q_x(1-q_x)}{N_x}} \qquad (5)$$

כאשר N_x הוא גודל האוכלוסייה בשכבת גיל x.

בתרשימים 2 ו-3 מוצגות פונקצית q_x ופונקציה לפי השיטה של H-P כאשר הפרמטרים נאמדו בשיטת ריבועים פחותים (Least squares estimator).

שני התרשימים מציגים הסתברויות למות של זכרים בגילים 15-30 באוכלוסייה היהודית. בתרשים 2 מוצגת התנהגות הפונקציה H-P (קו-"unab") והתנהגות פונקצית $y_x = log(q_x/(1-q_x))$. כמו כן, בתרשים זה ותרשימים הבאים מוצג רווח סמך עבור y_x (y_{max}, y_{min}). תרשים 2 מציג חוסר התאמה של הפונקציה בשיטת H-P לגילאי הצבא בישראל (גיל 19-20). פונקציה זו אינה מובהקת לפי רווח סמך של 95%. בישראל ישנה תמותה גבוהה בגילים אלה, ואילו אמידת הפונקציה בשיטת H-P מנמיכה את ההסתברויות למות בטווח גילים זה. לעומת זאת, בתרשים 3, בו ישנה השוואה בין פונקצית y_x ו-$\hat{y}_x$ מנובא ע"י רגרסיה (קו-"ypred"), ניתן לראות שפונקצית $\hat{y}_x$ נמצאת תמיד בתוך רווח הסמך של ההסתברות האמפירית.

תרשימים 4-5 מציגים את אי ההתאמה של הפונקציה בשיטת H-P לנתונים הישראלים והתאמה המבוססת על נוסחה (3) של רגרסיה.

תרשימים אלה מראים הסתברויות למות של נקבות בגילים 2-15 באוכלוסייה היהודית. בתרשים 4 ניתן לראות, כי הפונקציה של H-P נראית חריגה, היא יוצאת מגבולות רווח הסמך בגילים 6-11. כלומר הפונקציה מעלה את ההסתברות למות בגילים אלה. לעומת זאת, בתרשים 5 הפונקציה החדשה שנאמדה מבטאת את ההסתברויות בצורה מובהקת, תרשים הרגרסיה (קו-"ypred") לא יוצא מגבולות רווח הסמך בכל טווח הגילים.

בתרשימים 6-7 מוצגים הנתונים של הזכרים בגילים מבוגרים 80-92 באוכלוסייה הערבית בישראל. גם באוכלוסייה זו, הנוסחה של H-P (תרשים 6) מעלה את ההסתברות למות בגילים הגבוהים, ולעומת זאת הפונקציה החדשה (תרשים 7) עוברת באמצע רווח הסמך של ההסתברות למות.

בתרשימים 8-9 מוצגים הנתונים של זכרים (תרשים 8) ונקבות (תרשים 9) בגילים הצעירים באוכלוסייה הערבית. מהתרשימים ניתן לראות, שקו הרגרסיה רגיש יחסית לאי יציבות של הנתונים האמפיריים. בקרב זכרים ערבים ישנה אי סדירות מקרית בפונקצית התמותה בגיל 10 ובקרב נקבות ערביות בגיל 13. בגילים צעירים אלו יש מעט מאוד פטירות ולכן הנתון האמפירי אינו יציב. חשוב ששיטת ההחלקה לא תיתן משקל גבוה לאי סדירויות מקריות, אלא תחליק את הפונקציה. אלא ששיטת החישוב החדשה מחליקה רק במקצת את האי סדירויות, אך עדיין ישנם ביטויים של אי היציבות בפונקצית ההחלקה. מכאן, שבעתיד

צריכים לשפר את שיטת ההחלקה המוצעת במסמך זה, כך שתתן משקל קטן יותר בגילים בהם יש מעט מאוד פטירות.

בספרות נעשה שימוש בקריטריון אמידה נוסף של איבר דטרמיניסטי בפונקציה $\hat{y}_x$ וערך אמפירי y_x - הערך הממוצע של השגיאה היחסית (The mean magnitude of relative error)

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}.$$

לוח 1: השוואה בין MRE ל- $\hat{y}_i$ מ-(3) לבין MRE ל- $\hat{y}_i$ מאמידת H-P באמצעות תוכנת Mortpak

קבוצת אוכלוסייה	MRE-regression	MRE-Mortpak	הפרש %
זכרים			
סה"כ	0.0120	0.0163	*26.5*
יהודים ואחרים	0.0127	0.0200	*36.7*
יהודים	0.0132	0.0199	*33.3*
ערבים	0.0166	0.0331	*49.7*
נקבות			
סה"כ	0.0102	0.0200	*49.1*
יהודים ואחרים	0.0117	0.0182	*35.5*
יהודים	0.0118	0.0210	*43.6*
ערבים	0.0228	0.0299	*23.6*

לוח זה מציג, כי סכום ההפרשים הריבועיים בין y_x אמפירי לעומת $\hat{y}_x$ המנובא מהרגרסיה קטנים בכל קבוצות האוכלוסייה מסכום ההפרשים הריבועיים בין y_x האמפירי לעומת $\hat{y}_x$ מנוסחת H-P.

ה- MRE מהרגרסיה יותר טוב מה- MRE של H-P בממוצע ב- 36% בקרב זכרים ובכ-38% בקרב נקבות.

שימוש בפונקציות רציפות של לוח התמותה ובשיטות החישוב המפורטות בנספח א' מאפשר קבלת תוחלת חיים המוצגת בלוח הבא:

לוח 2:

תוחלת חיים לפי שיטות חישוב שונות, לפי קבוצת אוכלוסייה ומין ,שנים 1996-2000

קבוצת אוכלוסייה	לפני החלקה נוסחאות Chiang	שיטה חדשה	תוכנת Mortpak: לוח שלם בשיטת H-P UNABR	תוכנת Mortpak: לוח מקוצר LIFTB
זכרים				
סה"כ	76.2	76.3	76.4	76.3
יהודים ואחרים	76.6	76.6	76.8	76.7
יהודים	76.7	76.8	76.9	76.8
אוכלוסייה ערבית	74.2	74.3	74.4	74.7
נקבות				
סה"כ	80.3	80.3	80.6	80.3
יהודים ואחרים	80.7	80.7	80.9	80.7
יהודים	80.7	80.7	80.4	80.7
אוכלוסייה ערבית	77.3	77.4	77.3	77.3

מלוח 2 נראה, כי התוצאה החריגה ביותר היא לפי חישוב תוחלת החיים לגיל בודד לפי נוסחת H-P (שלם Mortpak). לפי שיטה זו, תוחלת החיים המתקבלת היא לרוב גבוהה יותר ורק בקבוצה אחת (יהודים/נקבה) היא נמוכה יותר. תוצאה זו נובעת מהחלקת יתר של הנתונים האמפיריים ע״י פונקצית H-P (ראה תרשימים 2,4,6), והחלקה זו בהרבה מקרים לא מובהקת סטטיסטית. תוחלת החיים המתקבלת על ידי השיטה המוצעת כאן זהה ברוב קבוצות האוכלוסייה לתוחלת חיים המחושבת לפי נתונים לפני ההחלקה בנוסחאות המקובלות. יחד עם זאת, השיטה החדשה מחליקה את ההסתברויות ומתקבלים נתונים יציבים יותר.

4. סיכום

שיטת ההחלקה החדשה שפותחה, מחליקה את הפונקציה של ההסתברויות למות ע״י פולינום דו-שלבי, בשימוש בשיטת אומדנים של נקודת שבר ושיטת נראות מקסימלית מקומית. השיטה מחלקת את טווח הנתונים לשני חלקים, עד לנקודת השבר ואחריה ובונה מקדמי רגרסיה שונים לשני טווחי גיל שונים. לפי המקדמים המתקבלים נבנים אומדנים עבור q_x, כאשר על סמך אומדנים אלו נבנה כל לוח התמותה.

התוצאה היא לוח תמותה עם תוחלת חיים שלא שונה באופן מובהק מתוחלת החיים המחושבת על סמך נתונים אמפיריים, ולא פחות חשוב המגמה של הסתברות למות יציבה יותר מקו אמפירי. היתרון המרכזי של השיטה נראה בהחלקה טובה יותר של הסתברויות למות של הזכרים היהודים בגיל 20-30. בתקופה זו נתוני התמותה של הזכרים שונים מנתונים של אוכלוסיות אחרות בגלל התמותה בצבא והשיטה החדשה מבטאת ייחודיות זו בהחלקה, כך שהנתונים המוחלקים בגילים אלו קרובים לנתונים האמפיריים ולא כמו בשיטה של Heligman-Pollard, שם רואים פער גדול בין הנתון האמפירי והנתון המנובא.

השיטה נבדקה ונמצאה כמתאימה לאמידת פונקצית התמותה באוכלוסייה הישראלית. לא בדקנו איך מתנהגת הפונקציה באוכלוסיות אחרות, לכן בעתיד חשוב לבדוק התנהגות המודל גם על נתונים של אוכלוסיות אחרות ולראות אם הוא אוניברסלי או שזהו מודל ייחודי לאוכלוסייה הישראלית בלבד.

שיטת ההחלקה המתוארת כאן יושמה כבר בפרסומי הלמ״ס (2003, 2004, 2005).

רשימת ספרות

Chiang, C.L. (1984) The Life Table and Its applications.

Fan, J., Farmen, M. and Gijbels, I. (1998) Local maximum likelihood estimation and inference. *J.R. Statist. Soc*. B. **60:** 591-608.

Hartmann, M. (1987) Past and Recent Attempts to Model Mortality at All Ages. *Journal of Official Statistics* **3:** 19-36.

Heligman, L. and Pollard, J.H. (1980) The Age pattern of Mortality. *Journal of the Institute of Actuaries* **107:** 49-75.

Kim H.J. and Siegmund D. (1989) The Likelihood Ratio Test for a Change-Point in Simple Linear Regression. *Biometrika* 76: 409-423.

Koul, H.L., Lianfen, Q. and Surgailis, D. (2003) Asymptotics of M-estimators in two-phase linear regression models. *Stochastic Processes and their Applications* **103:** 123-154.

Kostaki, A. (1991) The Heligman-Pollard Formula as a Tool for Expanding an Abridged Life Table. *Journal of Official Statistics* **7:** 311-323.

Kostaki, A. (1992) A Nine-Parameter Version of the Heligman-Pollard Formula. *Mathematical Population Studies* **3:** 277-288.

MORTPAK, The United Nations' Software Package for Mortality Measurement, United Nations, 1988.

Preston, S.H., Heuveline, P. and Guillot, M. (2001) Demography Measuring and Modeling Population Processes. *The life table and single decrement processes:* 37-70.

לוחות תמותה שלמים של ישראל 1996-2000, לקט ממצאים סטטיסטיים 2003/16.

לוחות תמותה שלמים של ישראל 1997-2001 1998-2002, לקט ממצאים סטטיסטיים 2004/12.

לוחות תמותה שלמים של ישראל 1999-2003, לקט ממצאים סטטיסטיים 2005/15.

נספח א'

מרכיבי לוח התמותה כפונקציות רציפות – Preston et al. (2001)

x = גיל מדויק x.

$l(x)$ = מספר אנשים שנשארו בחיים בגיל x.

${}_nd_x$ = מספר אנשים שנפטרו בין גיל x ל- $x+n$.

${}_nq_x$ = ההסתברות למות בין גיל x לגיל $x+n$.

${}_np_x$ = הסתברות לשרוד מגיל x עד גיל $x+n$.

${}_nL_x$ = מספר שנות אדם (person years) שחי הדור בין גיל x לגיל $x+n$.

T_x = מספר שנות אדם (person years) שנותרו אחרי גיל x.

e_x^0 = תוחלת חיים בגיל x.

${}_nm_x$ = שיעורי תמותה ממוצעים בין גיל x לגיל $x+n$.

${}_na_x$ = ממוצע מספר שנות אדם של אנשים שנפטרו בין גיל x לגיל $x+n$.

$$l(x) = l(a)e^{-\int_a^x \mu(y)dy} \quad for\ x > a$$

$${}_np_x = e^{-\int_x^{x+n}\mu(a)da}$$

$${}_nd_x = \int_x^{x+n} l(a)\mu(a)da = l(x)\int_x^{x+n} e^{-\int_x^a \mu(y)dy}\mu(a)da$$

$${}_nq_x = \int_x^{x+n} e^{-\int_x^a \mu(y)dy}\mu(a)da$$

$${}_nL_x = \int_x^{x+n} l(a)da = l(x)\int_x^{x+n} e^{-\int_x^a \mu(y)dy}da$$

$${}_nm_x = \frac{\int_x^{x+n} l(a)\mu(a)da}{\int_x^{x+n} l(a)da} = \frac{\int_x^{x+n} e^{-\int_x^a \mu(y)dy}\mu(a)da}{\int_x^{x+n} e^{-\int_x^a \mu(y)}da}$$

$${}_na_x = \frac{\int_x^{x+n} l(a)\mu(a)(a-x)da}{\int_x^{x+n} l(a)\mu(a)da} = \frac{\int_x^{x+n} e^{-\int_x^a \mu(y)dy}\mu(a)(a-x)da}{\int_x^{x+n} e^{-\int_x^a \mu(y)dy}\mu(a)da}$$

$$T_x = \int_x^{\infty} l(a)da = l(x)\int_x^{\infty} e^{-\int_x^a \mu(y)dy}da$$

$$e_x^0 = \frac{\int_x^\infty l(a)da}{l(x)} = \int_x^\infty e^{-\int_x^a \mu(y)dy} da = \frac{\int_x^\infty l(a)\mu(a)(a-x)da}{\int_x^\infty l(a)\mu(a)da}$$

מרכיבי לוח התמותה בחישוב בדיד – Chiang (1984)

$${}_n p_x = l(x+n)/l(x)$$

$${}_n d_x = l(x) - l(x+n)$$

$${}_n q_x = 1 - {}_n p_x$$

$${}_n m_x = {}_n d_x / {}_n L_x$$

$${}_n a_x = \left({}_n L_x - nl(x+n)\right)/{}_n d_x$$

$$e_x^0 = T_x / l_x$$

שיטות חישוב תוחלת חיים

לוח תמותה מבוסס על שיעורי הפטירה הסגוליים לפי גיל ומין, והוא מורכב מהפונקציות הבאות:

m_x - שיעורי התמותה בגיל x, כלומר מספר הנפטרים בגיל x מחולק באוכלוסייה הממוצעת בגיל x.

ערכי m_x לחישוב לוח התמותה לשנים 1996- 2000 מבוססים על שיעורי תמותה ממוצעים לשנים 1996-2000.

q_x - ההסתברות למות בין גיל x לגיל x+1, הטור מציג את חלקם היחסי של אלה שנפטרו בין גיל x לגיל x+1 מתוך אלה שנשארו בחיים עד גיל x.

${}_n a_x$ - ממוצע מספר שנות אדם של אנשים שנפטרו בין גיל x לגיל $x+n$.

l_x - מספר הנשארים בחיים בגיל מדויק x מתוך 100,000 נולדים (l_0 = 100,000 = שורש הלוח). ערכי l_x מחושבים על סמך ערכי q_x המתייחסים למספר הנשארים בחיים בגיל $x-1$.

$$l_x = l_{x-1}(1 - q_{x-1})$$

L_x - מספר שנות אדם (person years) שחי הדור שהגיע לגיל x, בין גיל x לגיל $x+1$.

$$L_x = l_{x+1} + a_x d_x$$

כאשר d_x - מספר אנשים שנפטרו בין גיל x ל- $x+1$.

$$d_x = l_x - l_{x+1}$$

מכיוון ששיטת החישוב מתבססת על גיל בודד, אנחנו מניחים כי $a_x \approx ½$ במרבית הגילים.

מכאן, $$L_x = (l_x + l_{x+1})/2$$

L_0 (מספר שנות אדם שחי כל הדור בין גיל 0 לגיל 1) ו- L_{100+} (מספר שנות אדם שחי כל הדור בגיל 100 ואילך) מחושבים בצורה שונה משתי סיבות:

L_0 מושפע מכך שהסתברות למות יורדת במהלך השנה הראשונה לחיים והקירוב של $a_x \approx ½$ אינו מתאים. כדי לבטא זאת אנחנו מניחים ש: $a_x = 0.3$.

$$L_0 = 0.3l_0 + 0.7l_1$$

L_{100+} מבטא את הצורך להעריך את יתרת שנות החיים שיחיה הדור עד שימות האחרון ממנו.

$$L_{100+} = 1000(l_{100}/m_{100+})$$

T_x - סך כל שנות אדם שנותרו לשורדי הדור לחיות לאחר הגיעם לגיל x. T_x מתקבל כסכום של L_x עבור כל הגילים הגבוהים מ- x.

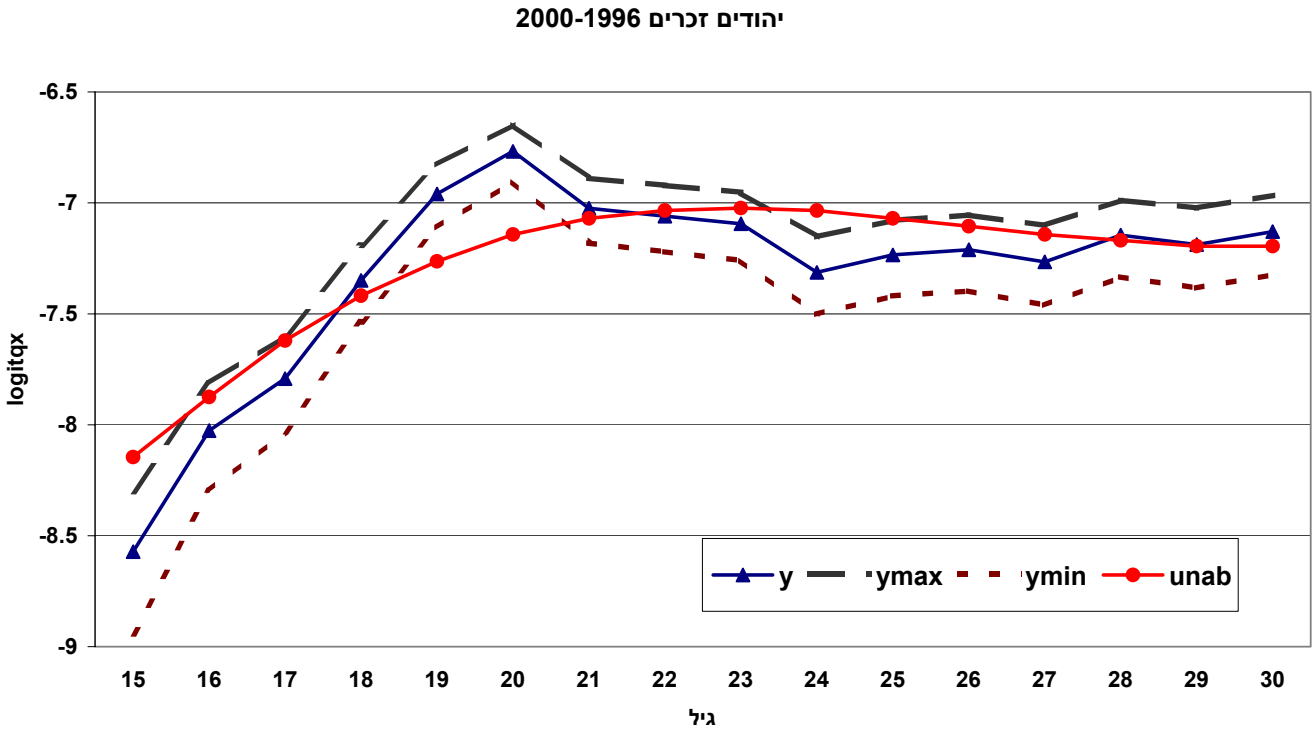
ערכי q_x עבור הגילים משנה אחת ומעלה נגזרים מערכי m_x לפי הנוסחה: $q_x = \frac{m_x}{1 + ½ m_x}$,

תחת ההנחה כי $a_x \approx ½$.

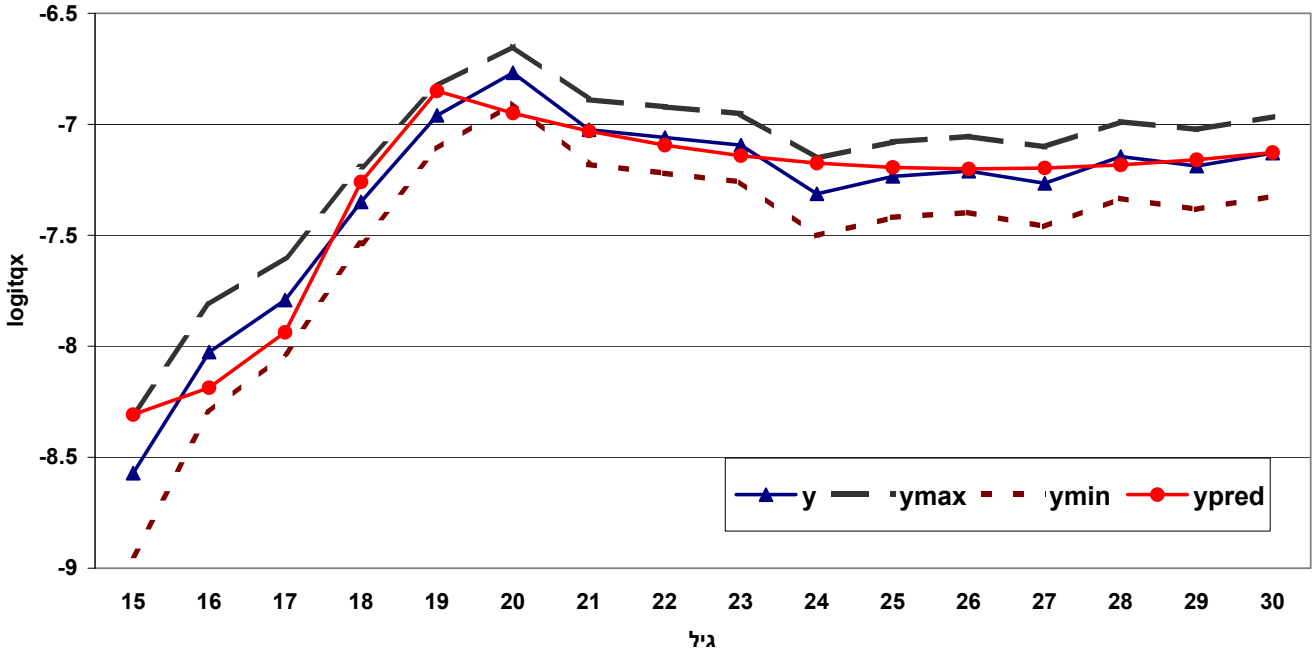
e_x^0 - תוחלת חיים בגיל x, הנה ממוצע שנות החיים שנותרו לאדם לחיות בגיל x, בהנחה שנשאר בחיים עד לגיל x, ובהנחה שדפוסי התמותה נשארים קבועים.

$$e_x^0 = T_x/l_x$$

תרשים 2: השוואה בין הסתברות למות אמפירית (קו-"Y") לבין הסתברות המתקבלת מאמידת פונקצית Heligman-Pollard בתוכנת Mortpak (קו-"Unab")

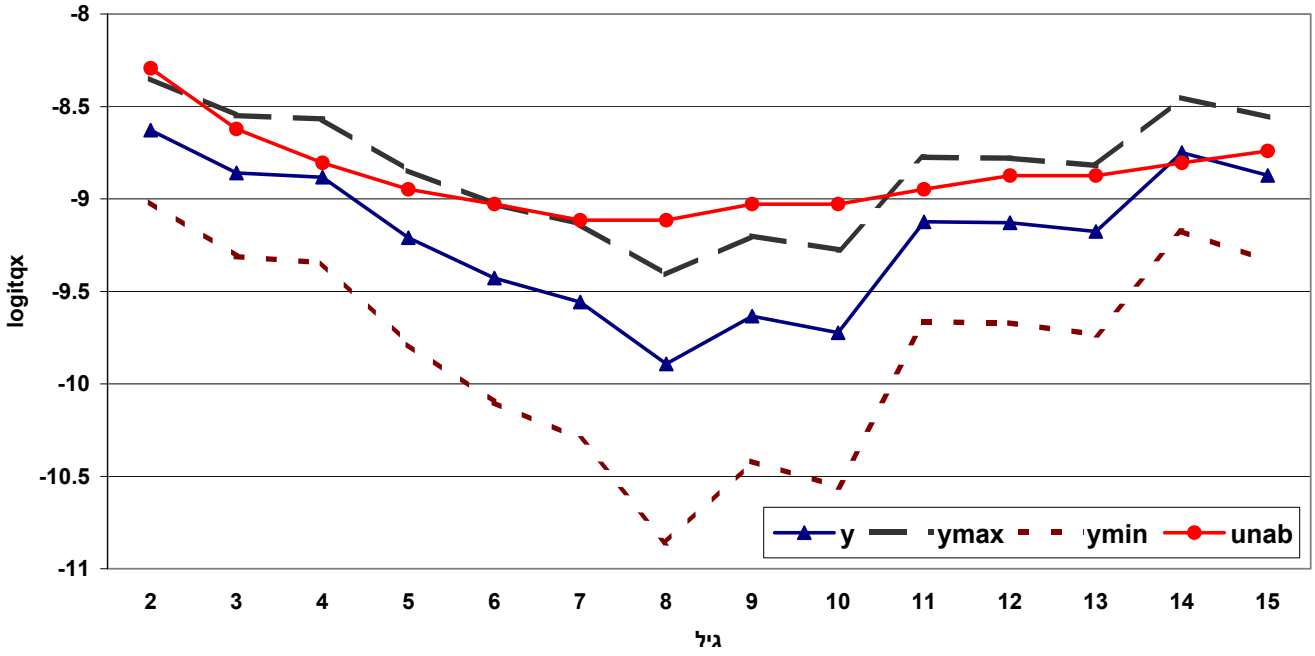


תרשים 3 : השוואה בין הסתברות למות אמפירית (קו-"Y") לבין הסתברות המתקבלת ממודל הרגרסיה (קו-"ypred")



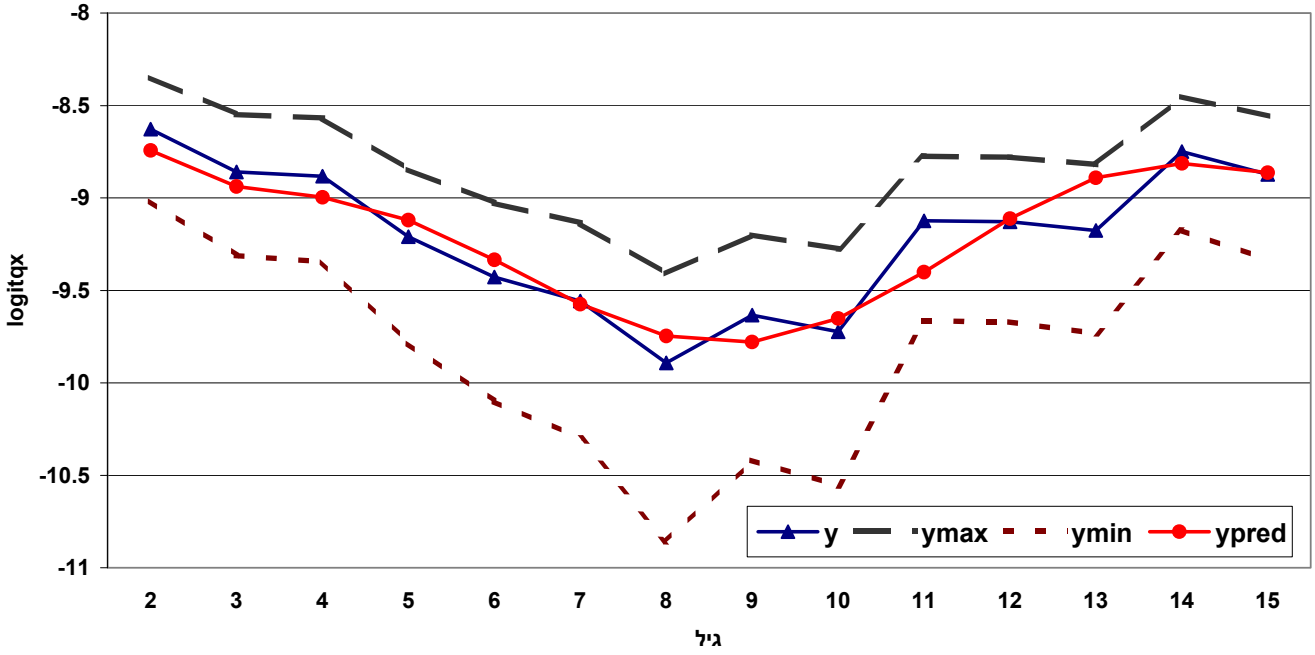
תרשים 4: השוואה בין הסתברות למות אמפירית (קו-"Y") לבין הסתברות המתקבלת מאמידת פונקצית Heligman-Pollard בתוכנת Mortpak (קו-"Unab")

יהודים נקבות גילים צעירים 1996-2000

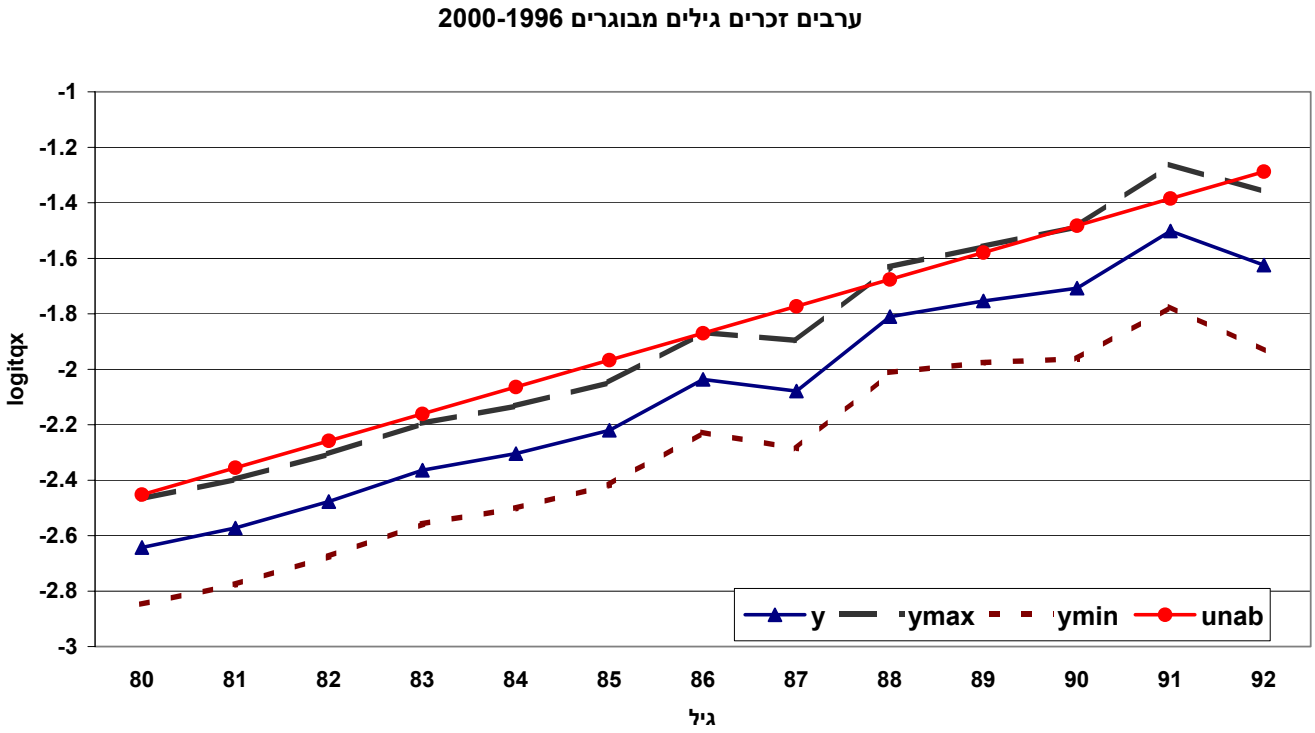


תרשים 5: השוואה בין הסתברות למות אמפירית (קו-"Y") לבין הסתברות המתקבלת ממודל הרגרסיה (קו-"ypred")

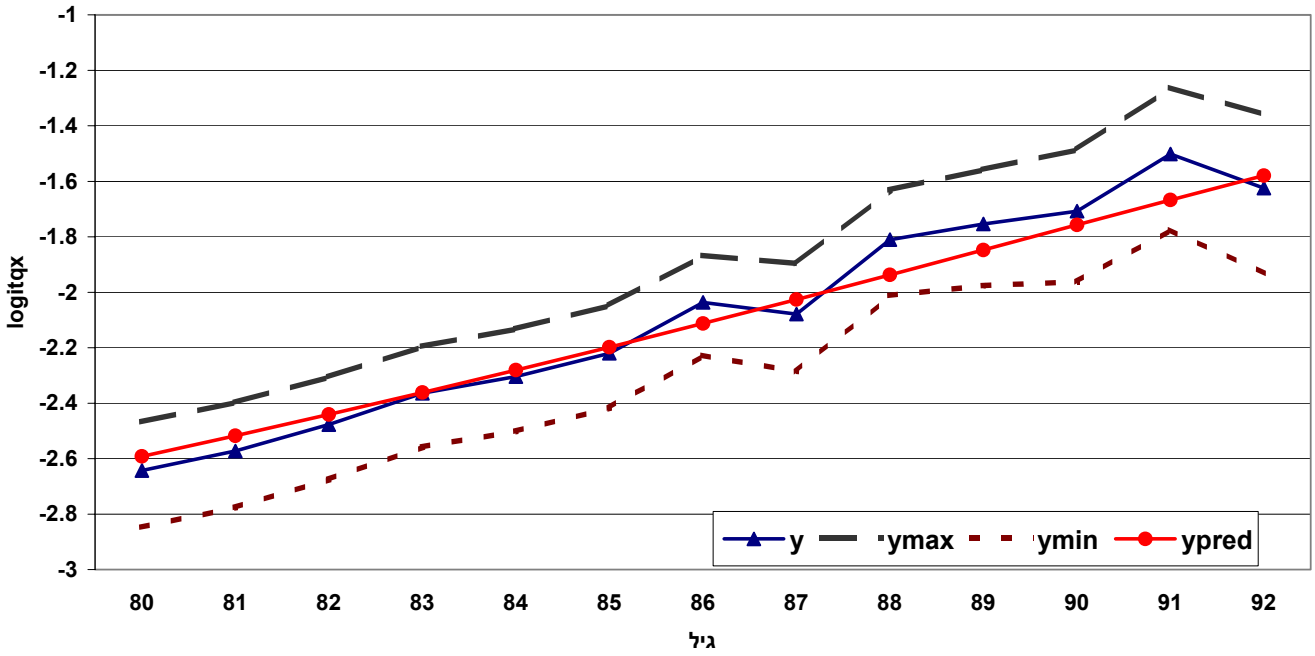
יהודים נקבות גילים צעירים 1996-2000



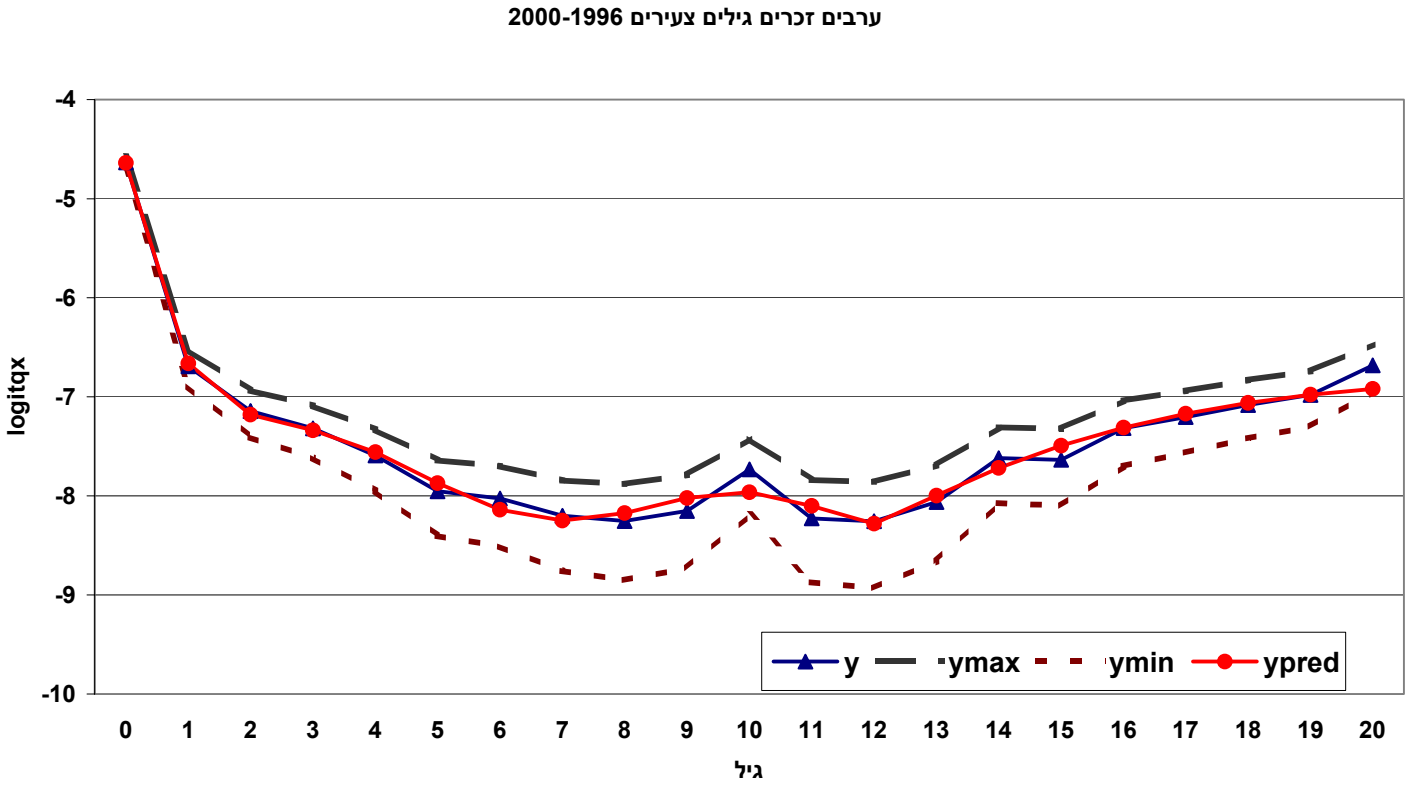
תרשים 6: השוואה בין הסתברות למות אמפירית (קו-"Y") לבין הסתברות המתקבלת מאמידת פונקצית Heligman-Pollard בתוכנת Mortpak (קו-"Unab")



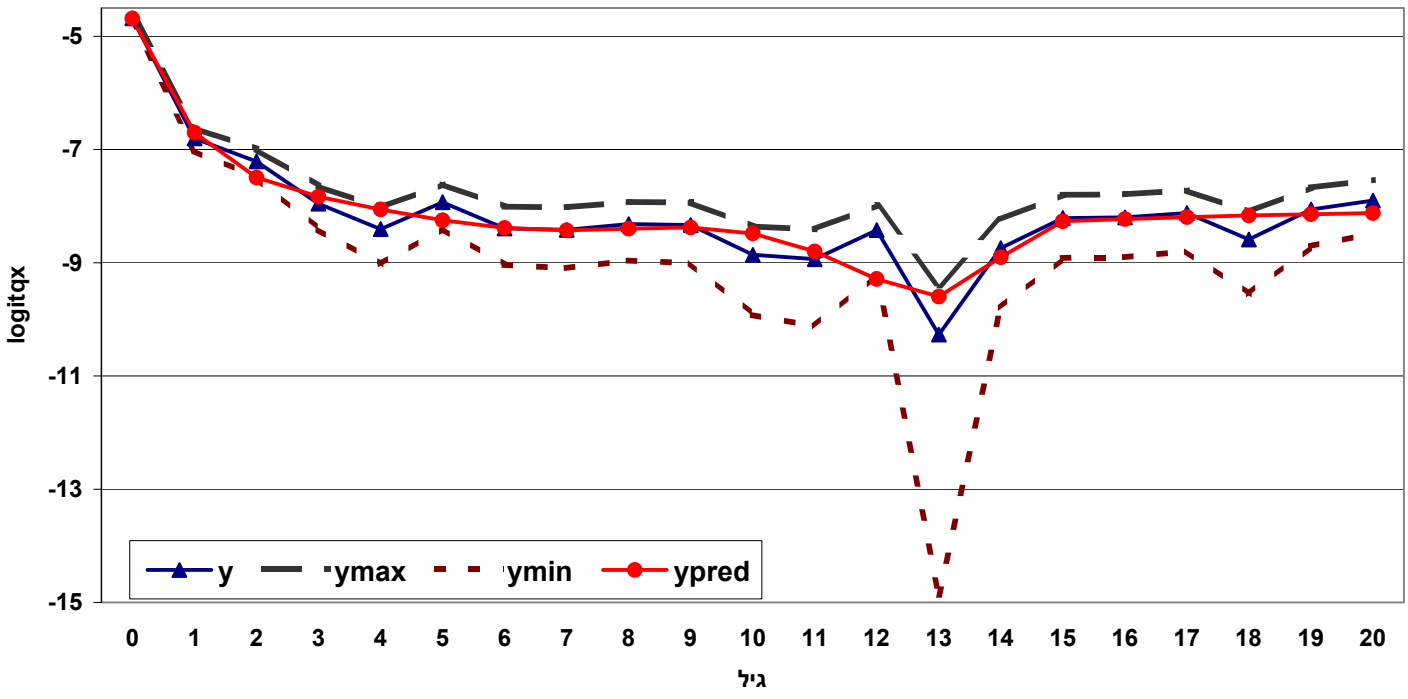
תרשים 7: השוואה בין הסתברות למות אמפירית (קו-"Y") לבין הסתברות המתקבלת ממודל הרגרסיה (קו-"ypred")



תרשים 8: השוואה בין הסתברות למות אמפירית (קו-"Y") לבין הסתברות המתקבלת ממודל הרגרסיה (קו-"ypred")

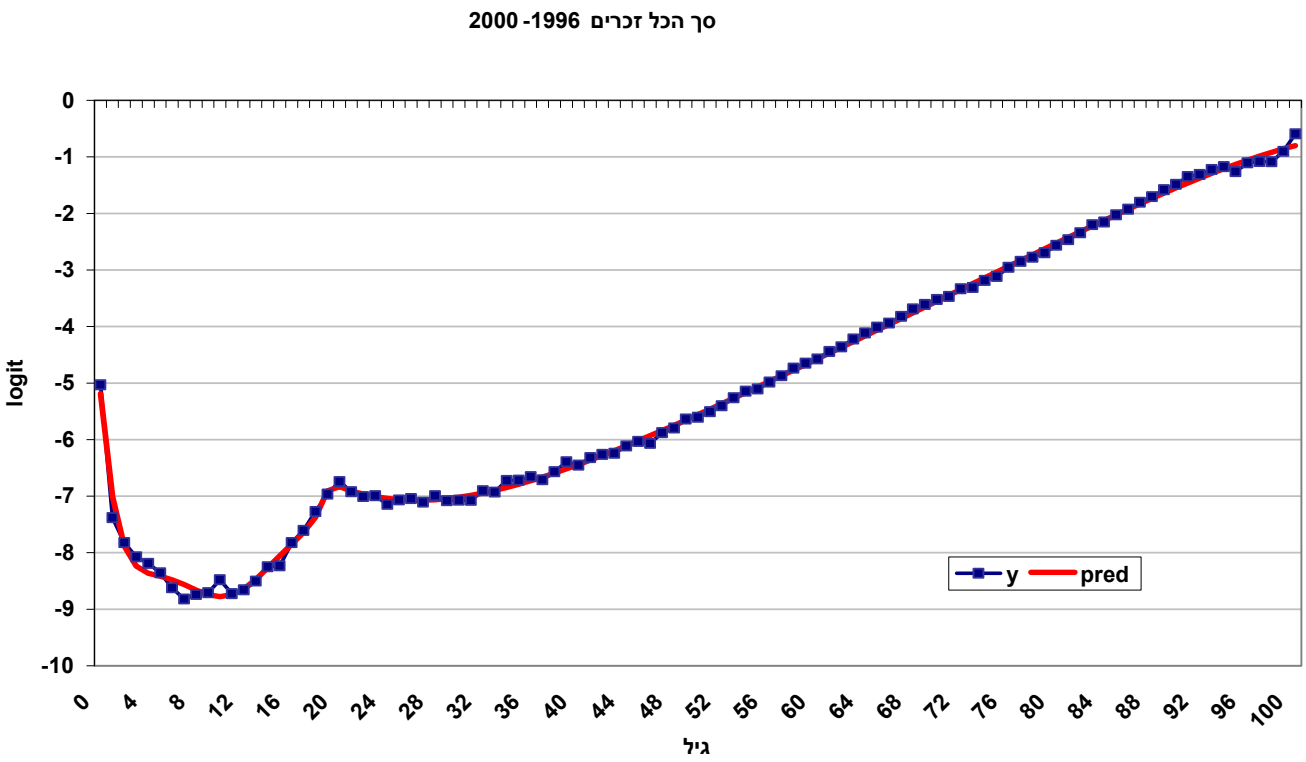


תרשים 9: השוואה בין הסתברות למות אמפירית (קו-"Y") לבין הסתברות המתקבלת ממודל הרגרסיה (קו-"ypred")

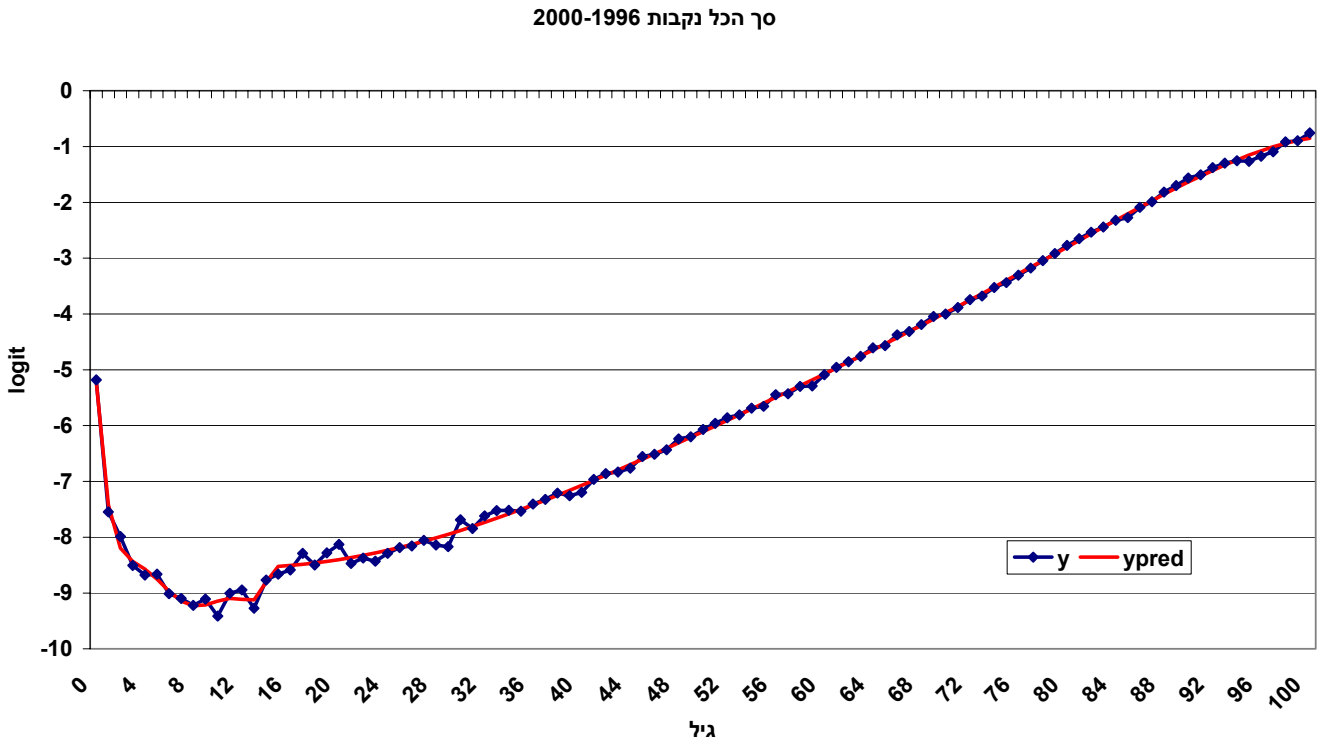


תרשימים 10-17 מציגים השוואה בין הסתברות למות אמפירית (קו-"Y") לבין הסתברות המתקבלת ממודל הרגרסיה (קו-"ypred") בכל קבוצות האוכלוסייה בגילים 0-100

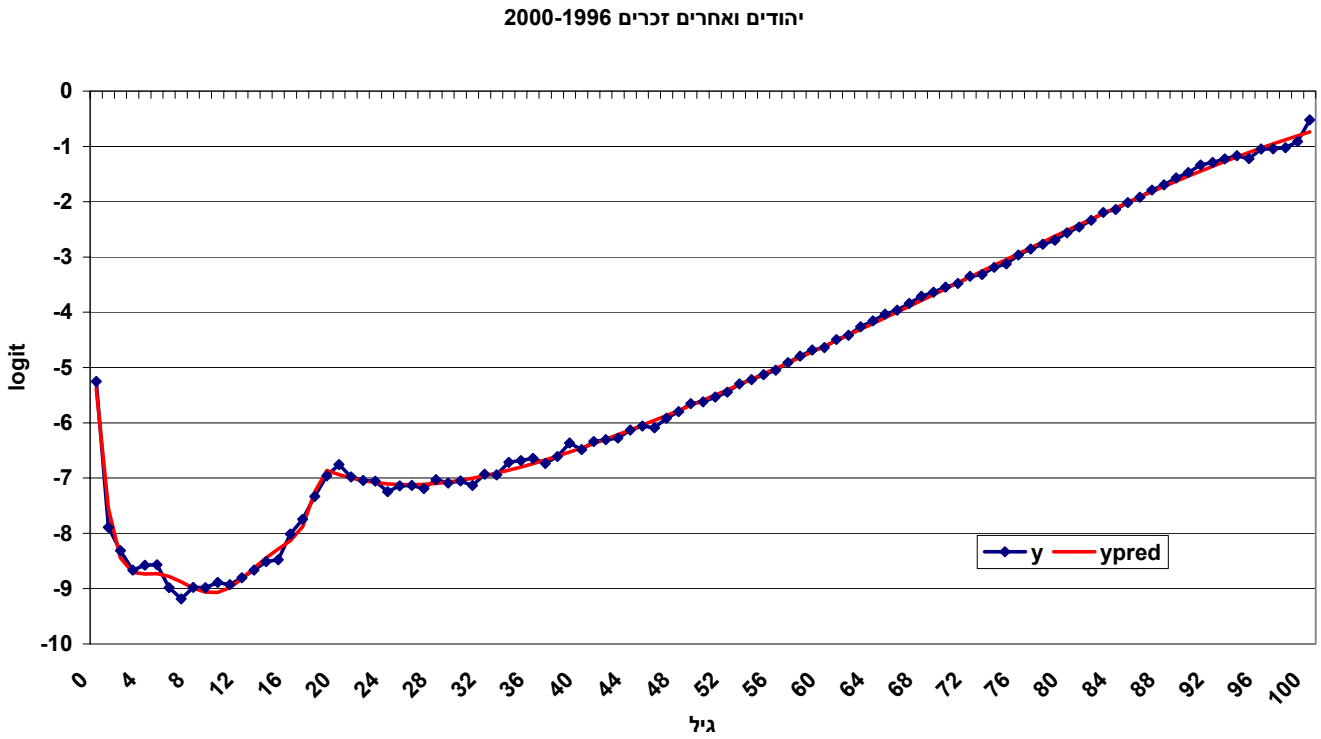
תרשים 10:



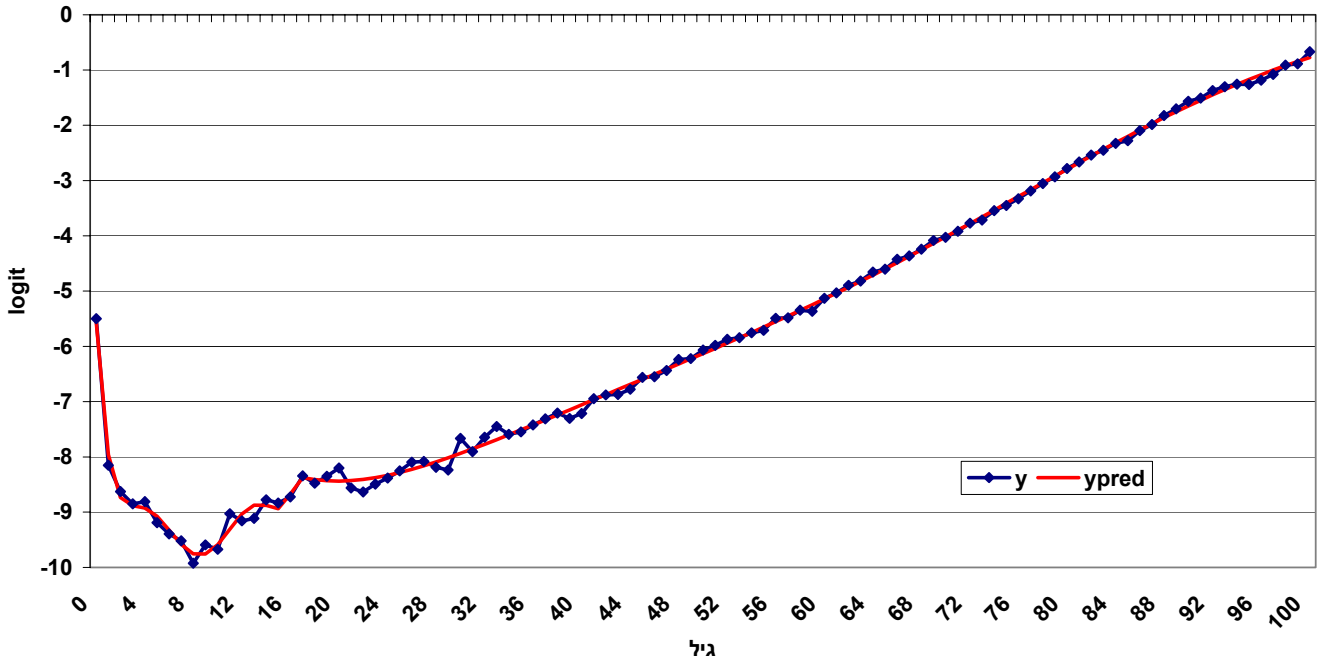
תרשים 11:



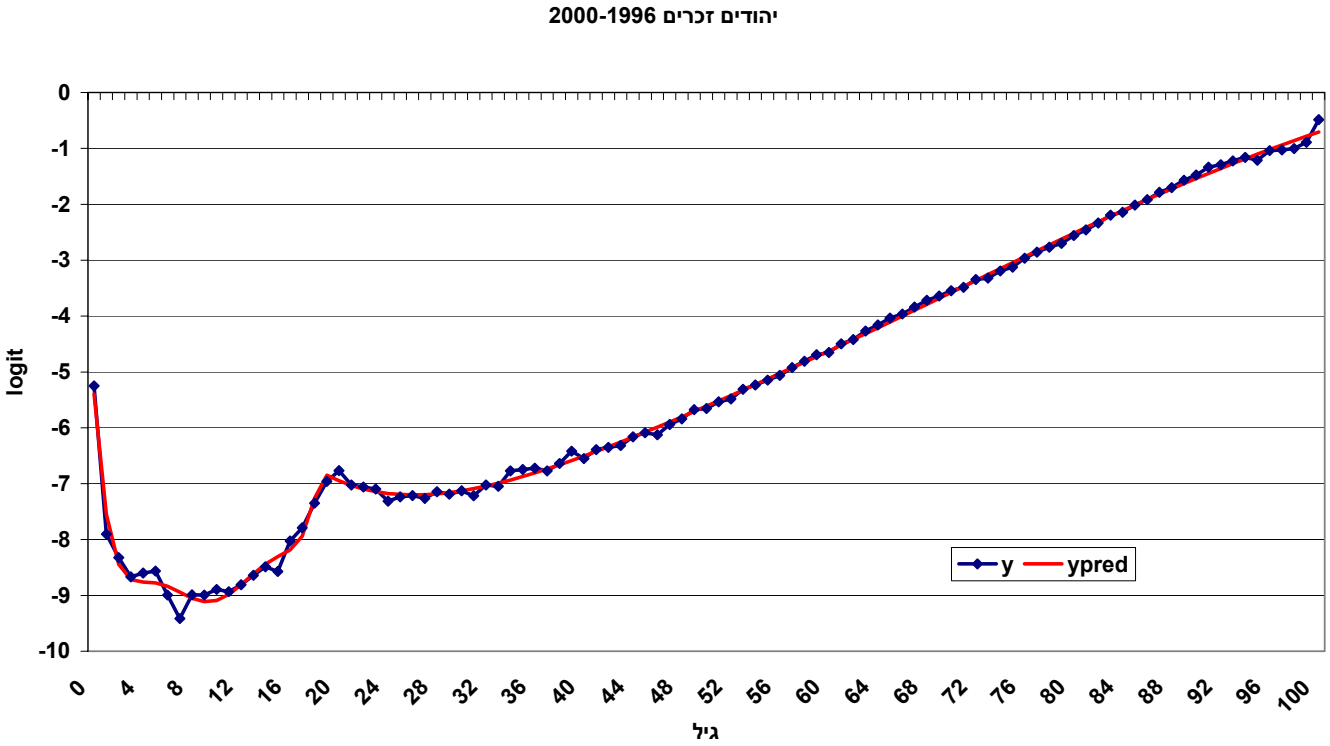
תרשים 12:



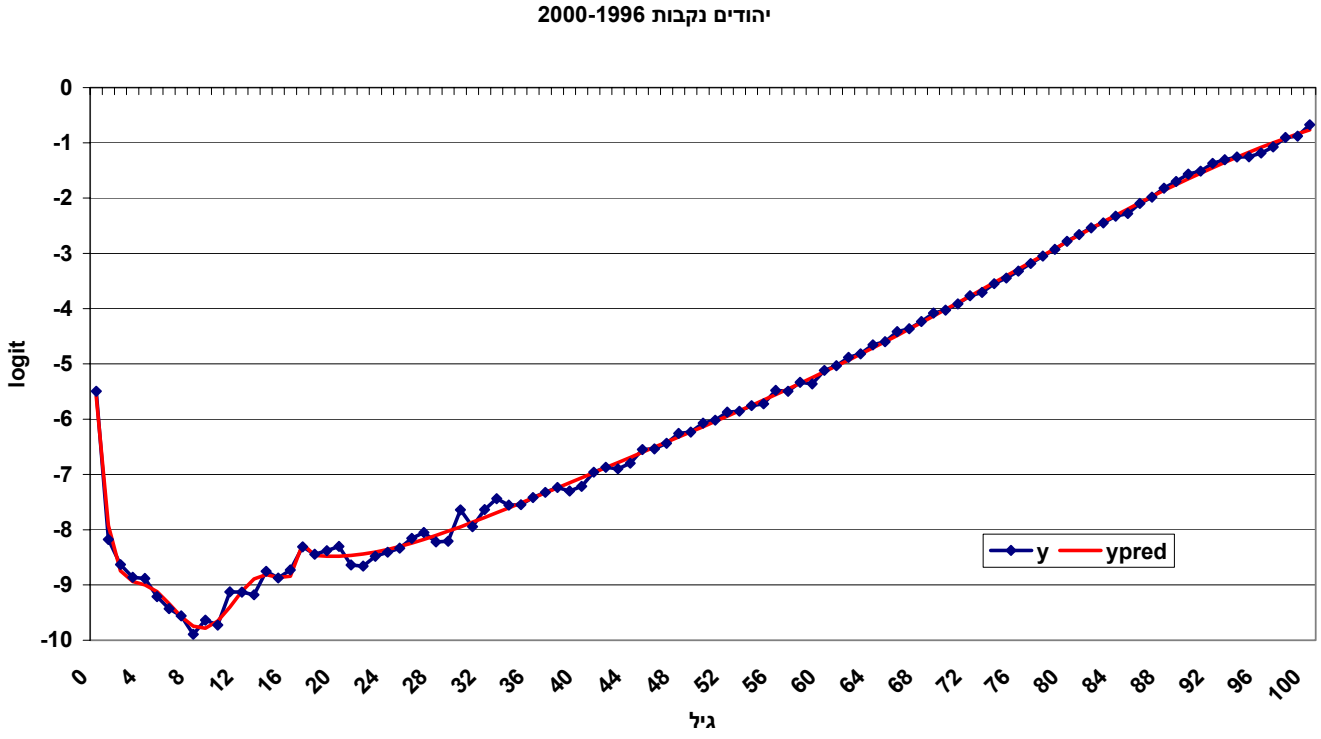
תרשים 13:



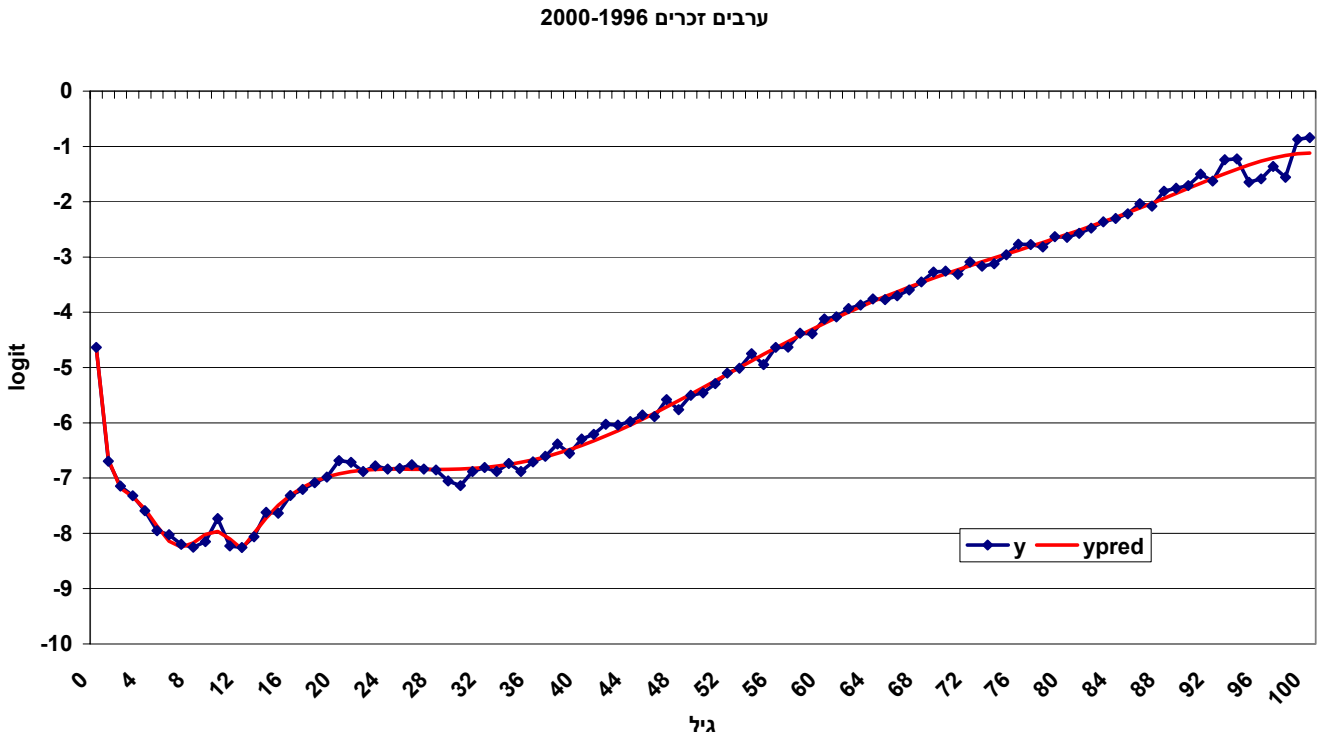
תרשים 14:



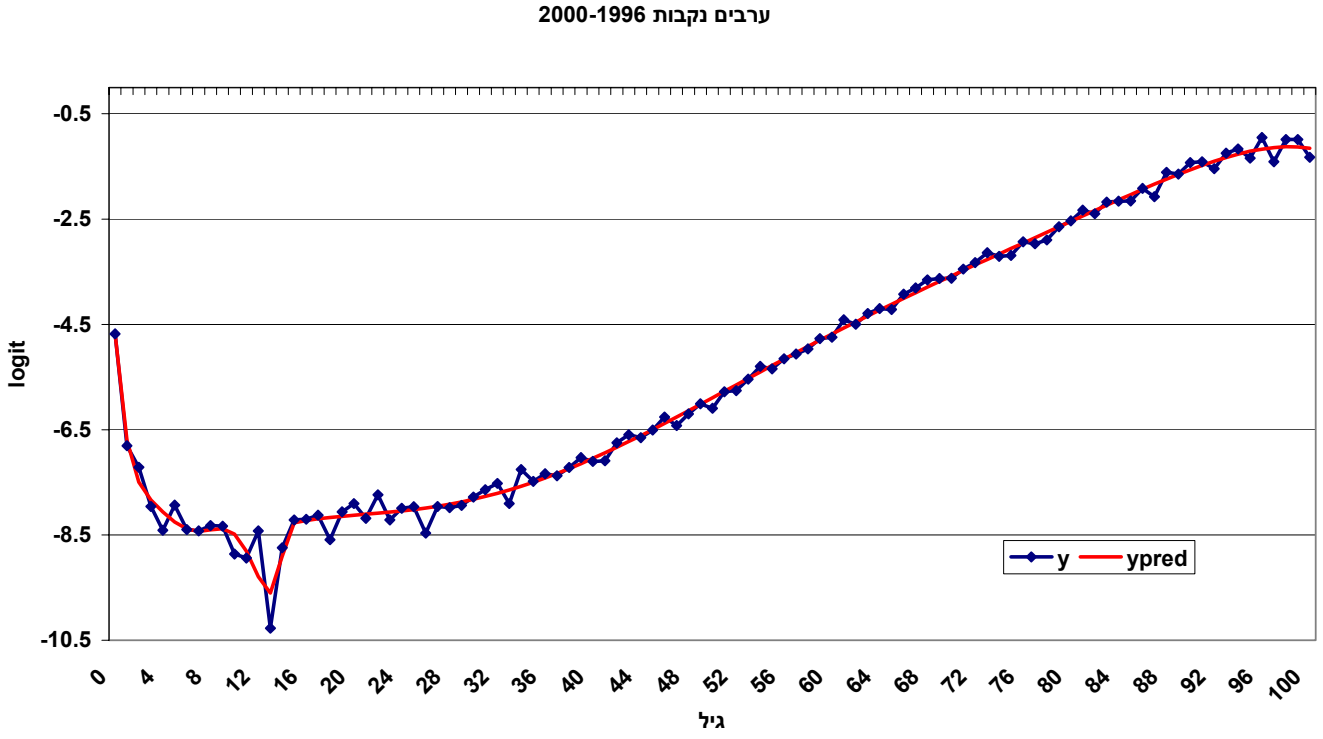
תרשים 15:



תרשים 16:



תרשים 17:



נספח ג'

מקדמי הרגרסיה ונקודות השבר								
קבוצת אוכלוסייה	סך הכל	יהודים ואחרים	יהודים	אוכלוסייה ערבית	סך הכל	יהודים ואחרים	יהודים	אוכלוסייה ערבית
מין	זכרים				נקבות			
Xc	20	20	20	13	15	18	18	15
x<Xc								
Intercept	-5.18516	-5.41886	-5.43260	-4.64030	-5.20799	-5.58832	-5.58865	-4.68578
x	-2.49266	-2.89788	-2.85793	-3.41460	-3.38191	-3.50668	-3.51819	-3.03581
x2	1.56731	1.92940	1.88840	3.63163	2.91519	2.89578	2.89545	2.57564
x3	-0.74592	-0.93925	-0.92227	-2.91787	-1.93382	-1.75518	-1.75174	-1.82218
x4	0.24407	0.31122	0.30836	1.56435	0.88009	0.70892	0.70676	0.92141
x5	-0.04893	-0.06307	-0.06321	-0.50383	-0.24574	-0.17183	-0.17120	-0.28898
x6	0.00459	0.00598	0.00606	0.07459	0.03204	0.01914	0.01907	0.04203
x≥Xc								
Intercept	-0.35373	1.80452	4.23644	-20.66411	-8.49445	-2.88731	-2.54262	-12.00667
x	-0.67687	-0.98731	-1.27897	1.99234	-0.00256	-0.77453	-0.83387	0.62143
x2	0.04857	0.08213	0.10861	-0.22240	-0.00323	0.07777	0.08475	-0.07877
x3	-0.00237	-0.00505	-0.00687	0.01820	0.00091	-0.00540	-0.00599	0.00731
x4	0.00008	0.00023	0.00032	-0.00102	-0.00008	0.00027	0.00030	-0.00044
x5	0.00000	-0.00001	-0.00001	0.00004	0.00000	-0.00001	-0.00001	0.00002
x6	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

$$x_6 = \frac{x^6}{6!} \quad x_5 = \frac{x^5}{5!} \quad x_4 = \frac{x^4}{4!} \quad x_3 = \frac{x^3}{3!} \quad x_2 = \frac{x^2}{2}$$