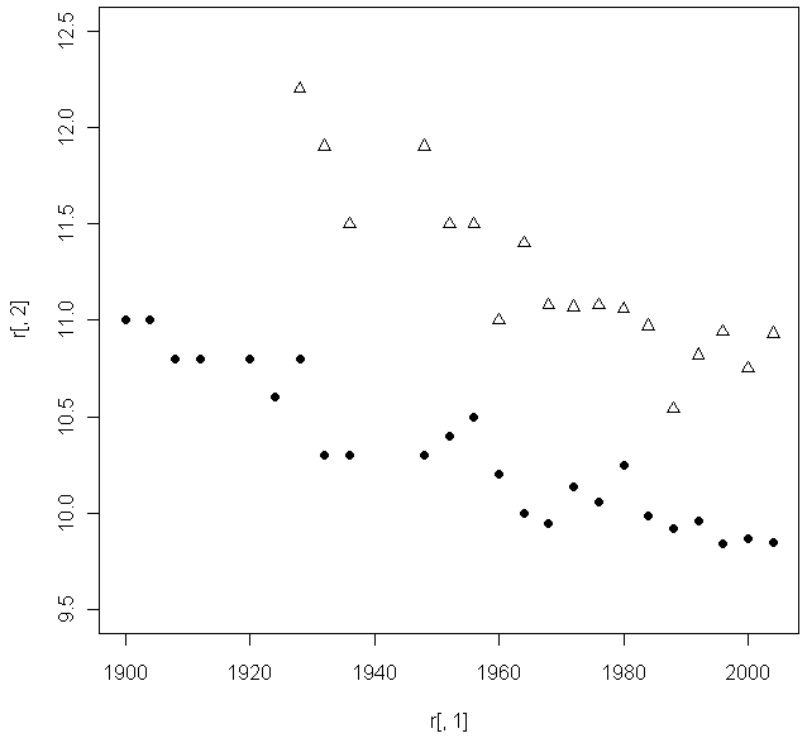
סטטיסטיקה / תרגיל #3

אריאל סטולרמן
קבוצה 03

(1)

(i)

```
> r=read.csv("e:\\sprinttimes.csv")
>  plot(r[,2]~r[,1],  pch=19,  xlim=c(1900,2004),  ylim=c(9.5,12.5));points(r[,3]~r[,1],
pch=24)
```



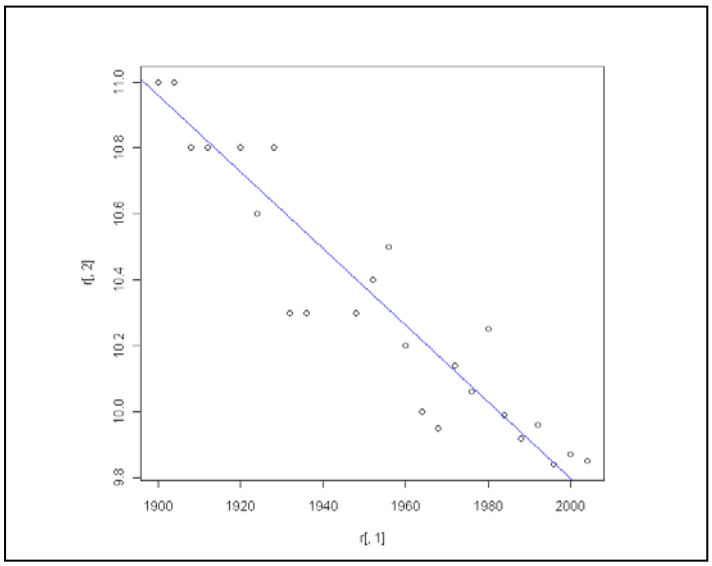
נראה כי קשר לינארי בין השנה לזמני הריצה סביר לאור התרשים – לנשים ולגברים בנפרד כמובן. ניתן ללמוד מזה שתוצאות זמני הריצה (עבור גברים ועבור נשים בנפרד) יורדות מקצב כמעט קבוע – כלומר ישנו שיפור קבוע לאורך הזמן. כמו כן ניתן לראות בתרשים כי קצב שיפור הנשים גדול מקצב שיפור הגברים, וכי הגברים (עד כה) רצים מהר מנשים.

(ii)

<u>הקו החסין של תוצאות הגברים:</u>

```
> rline(r[,2],r[,1])
$intercept
[1] 33.04203

$slope
[1] -0.01162162

$pred
 [1] 10.960946 10.914459 10.867973 10.821486 10.728514 10.682027 10.635541
 [8] 10.589054 10.542568 10.403108 10.356622 10.310135 10.263649 10.217162
```

```
[15] 10.170676 10.124189 10.077703 10.031216  9.984730  9.938243  9.891757
[22]  9.845270  9.798784  9.752297

$resid
 [1]  0.03905405  0.08554054 -0.06797297 -0.02148649  0.07148649 -0.08202703
 [7]  0.16445946 -0.28905405 -0.24256757 -0.10310811  0.04337838  0.18986486
[13] -0.06364865 -0.21716216 -0.22067568  0.01581081 -0.01770270  0.21878378
[19]  0.00527027 -0.01824324  0.06824324 -0.00527027  0.07121622  0.09770270

$x
 [1] 1900 1904 1908 1912 1920 1924 1928 1932 1936 1948 1952 1956 1960 1964
[15] 1968 1972 1976 1980 1984 1988 1992 1996 2000 2004

$xb
[1] 1916

$yb
[1] 10.8

$xt
[1] 1990

$yt
[1] 9.94
```




```
> plot(r[,2]~r[,1]);abline(a=33.04203, b=-0.01162162, col='blue')
```

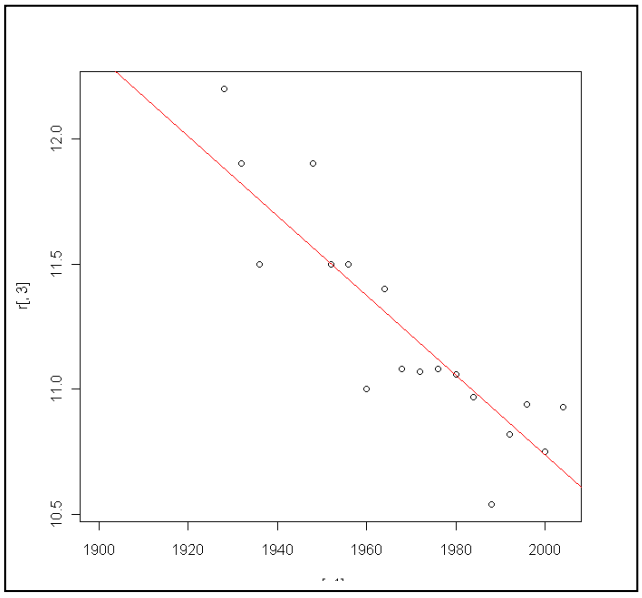
שיפוע הקו החסין של תוצאות הגברים הוא -0.01162162, והמשמעות היא שישנה ירידה בקצב קבוע (בערך) של זמני הריצה של הגברים לאורך ציר הזמן.

<u>הקו החסין של תוצאות הנשים:</u>

```
$intercept
[1] 42.47135

$slope
[1] -0.01586538

$pred
 [1] 11.88288 11.81942 11.75596 11.56558 11.50212 11.43865 11.37519 11.31173
 [9] 11.24827 11.18481 11.12135 11.05788 10.99442 10.93096 10.86750 10.80404
[17] 10.74058 10.67712

$resid
 [1]  0.317115385  0.080576923 -0.255961538  0.334423077 -0.002115385
 [6]  0.061346154 -0.375192308  0.088269231 -0.168269231 -0.114807692
[11] -0.041346154  0.002115385 -0.024423077 -0.390961538 -0.047500000
```

```
[16]  0.135961538  0.009423077  0.252884615
```

```
$x
 [1] 1928 1932 1936 1948 1952 1956 1960 1964 1968 1972 1976 1980 1984 1988
[15] 1992 1996 2000 2004

$xb
[1] 1942

$yb
[1] 11.7

$xt
[1] 1994

$yt
[1] 10.875
```



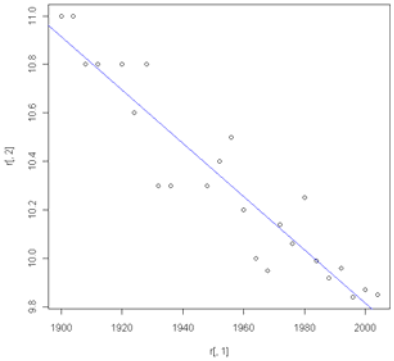

```
> plot(r[,3]~r[,1]);abline(a=42.47135, b=-0.01586538, col='red')
```

שיפוע הקו החסין של תוצאות הנשים הוא -0.01586538, והמשמעות היא שישנה ירידה בקצב קבוע (בערך) גם בזמני הריצה של הנשים לאורך ציר הזמן, והקצב מהיר יותר מאשר אצל הגברים (שיפוע חד יותר).

(iii)

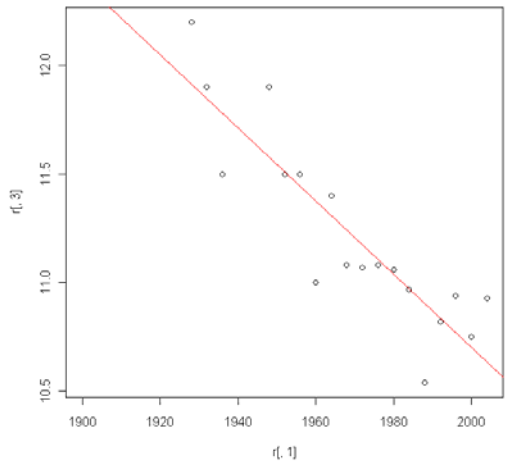
```
> r1=r[,1][!is.na(r[,2])]
> r2=r[,2][!is.na(r[,2])]
> ols.line(r1,r2)
$slope
[1] -0.01100556

$intercept
[1] 31.82645
```

```
> plot(r[,2]~r[,1]);abline(a=31.82645, b=-0.01100556, col='blue')
```




```
> r1=r[,1][!is.na(r[,3])]
```

```
> r3=r[,3][!is.na(r[,3])]
> ols.line(r1,r3)
$slope
[1] -0.01682207

$intercept
[1] 44.34705

> plot(r[,3]~r[,1]);abline(a=44.34705, b=-0.01682207, col='red')
```



שיפועי קווי הריבועים הפחותים:

- גברים: -0.01100556
- נשים: -0.01682207

שיפור הנשים מידי שנה גדול יותר לפי שיטת קו הריבועים הפחותים (בו השיפוע גדול יותר בערכו המוחלט, כלומר קצב שיפור גדול יותר).

(iv)

אני מסכים עם מסקנת הכותבים שלא ניתן לחזות את צמצום הפערים בין נשים וגברים, למרות מה שניתן לפרש מהנתונים עד כה, וזאת כיוון שישנם פרמטרים רבים אחרים המשפיעים על שיפור יכולות כגון אלו שהוצגו במאמר (שימוש בסמים, אקלים וכיו"ב), ולכן הסקת מסקנות רק על בסיס תוצאות אולימפיות לאורך השנים אינה מלאה ומשקפת דיו.

(v)

לפי הקו החסין:

$$33.04203 - 0.01162162x = 42.47135 - 0.01586538x \Leftrightarrow$$

$$0.00424376x = 9.42932 \Leftrightarrow$$

$$x = 2221.9258\sim$$

← לפי הקו החסין הנשים יעקפו את הגברים באולימפיאדה של שנת 2224.

לפי קו הריבועים הפחותים:

$31.82645 - 0.01100556x = 44.34705 - 0.01682207x \Leftrightarrow$

$0.00581651x = 12.5206 \Leftrightarrow$

$x = 2152.5966\sim$

← לפי קו הריבועים הפחותים הנשים יעקפו את הגברים באולימפיאדה של שנת 2156.

(2)

(a) להלן התאמת מישור ריבועים פחותים על הנתונים:

```
> dat=read.table("e:\\ex03data", header=T)
> lm(dat)

Call:
lm(formula = dat)

Coefficients:
(Intercept)      x1          x2
      1.924   0.965       3.740
```

(b) משוואת הניבוי היא משוואת המישור הבאה: $y = 1.924 + 0.965x_1 + 3.740x_2$

(c)

נשתמש בטרנספורמציה $\cos \circ \sin$ על x_1 וגם על x_2, כלומר משוואת הניבוי אחרי טרנספורמציה תהיה:

$y = 1.924 + 0.965*\cos(\sin(x_1)) + 3.740*\cos(\sin(x_2))$

עדות לכך שטרנספורמציה זו טובה מתבטאת ב-r^2.

(d)

השונות המוסברת לפני ואחרי הטרנספורמציות:

```
> dat=read.table("e:\\ex03data", header=T)
> y=dat$y
> x1=dat$x1
> x2=dat$x2
> y.hat=1.924 + 0.965*x1 + 3.740*x2
> y.hat.trans=1.924 + 0.965*cos(sin(x1)) + 3.740*cos(sin(x2))
> r.squared=function(observed,predicted){
+ ssr=sum((predicted-mean(observed))^2)
+ sst=sum((observed-mean(observed))^2)
+ ssr/sst
+ }
> r.squared(y,y.hat)
[1] 0.6759625
> r.squared(y,y.hat.trans)
[1] 0.8706616
```

השונות המוסברת לפני: 0.6759625

השונות המוסברת אחרי: 0.8706616

נראה כי הטרנספורמציות אכן עזרו להסביר את המחירים – יש קשר (השואף ל) לינארי חיובי בין מחירי החולצות ובין מדד הזמן ואחוז המודעות בציבור – ככל שעובר הזמן ומודעות הציבור גדלה, מחירי החולצות עולים.

(3)

(i)

A: **r יהיה קרוב ל-0** כיוון שישנו קשר שאינו לינארי בין הנתונים (קשר ריבועי). לא ניתן לקבוע את סימן r בבירור.

B: **r יהיה חיובי וקרוב מאוד ל-1** (ואפילו שווה ממש ל-1) כיוון שישנו קשר לינארי מובהק בין כל הנתונים – הם יושבים ממש על משוואת ישר כלשהו ששיפועו חיובי.

C: **r יהיה קרוב ל-0** כיוון שישנו קשר שאינו לינארי בין הנתונים (קשר מחזורי כלשהו כדוגמת sin). לא ניתן לקבוע את סימן r בבירור.

E: **r יהיה קרוב ל-0**.

F: **r יהיה קרוב ל-0** כיוון שלא נראה שקיים קשר בין x ל-y.

G: **r יהיה קרוב ל-1-** (אולי לא קרוב מידי) כיוון שנראה כי קיים קשר לינארי שלילי בין x ל-y.

(ii)

A: אחוז הפיזור של הנתונים המוסבר ע"י הקשר הלינארי במקרה זה **נמוך**, שכן אין קשר לינארי בין הנתונים (יש קשר שאינו לינארי).

B: אחוז הפיזור של הנתונים המוסבר ע"י הקשר הלינארי במקרה זה **גבוה** (ואף 100%), שכן יש קשר לינארי מובהק בין הנתונים.

C: אחוז הפיזור של הנתונים המוסבר ע"י הקשר הלינארי במקרה זה **נמוך**, שכן אין קשר לינארי בין הנתונים (יש קשר שאינו לינארי).

E: אחוז הפיזור של הנתונים המוסבר ע"י הקשר הלינארי במקרה זה **נמוך**, שכן אין קשר לינארי בין הנתונים.

F: אחוז הפיזור של הנתונים המוסבר ע"י הקשר הלינארי במקרה זה **נמוך**, שכן אין קשר לינארי בין הנתונים.

G: אחוז הפיזור של הנתונים המוסבר ע"י הקשר הלינארי במקרה זה **בינוני-גבוה**, שכן ישנו קשר לינארי בין הנתונים, אך הפיזור רב ולכן לא יהיה גבוה כמו ב-B.

(iii)

A: המתאם יתקרב מאוד ל-1, כיוון ש-y דומה ל-x^2 (הנחה), ולכן אם $x_{new}=x^2$ אז y כפונקציה של x החדש היא פונקציה לינארית עם שיפוע חיובי.

B: נניח שאכן מתקיים קשר לינארי, נסמן את משוואת הישר הנתון כ-y=ax+b ונבצע את הטרנספורמציה הבאה:

$y_{new}=f(y)=c+dy,\ x_{new}=f(x)=c+dx$ ➔

$y_{new}=ax_{new}+b \Leftrightarrow c+dy=a(c+dx)+b \Leftrightarrow dy = adx + ac + b - c \Leftrightarrow$

y = ax + (ac+b-c)/d

➔ הטרנספורמציה תשאיר את המשוואה לינארית עם אותו שיפוע, ורק תשנה את החיתוך עם ציר ה-y, ולפיכך מקדם המתאם ישאר חיובי וקרוב (אם לא שווה) ל-1.

G: הטרנספורמציה תהפוך את הקשר בין x ו-y מלינארי שלילי לפרבולי (שורש ריבועי) שלילי, ולכן תחת הנחה שהקשר הלינארי המקורי חזק מספיק (אחוז שונות מוסברת גבוה), נקבל כי כי מקדם המתאם יתקרב ל-0 על שום אובדן הקשר הלינארי.