

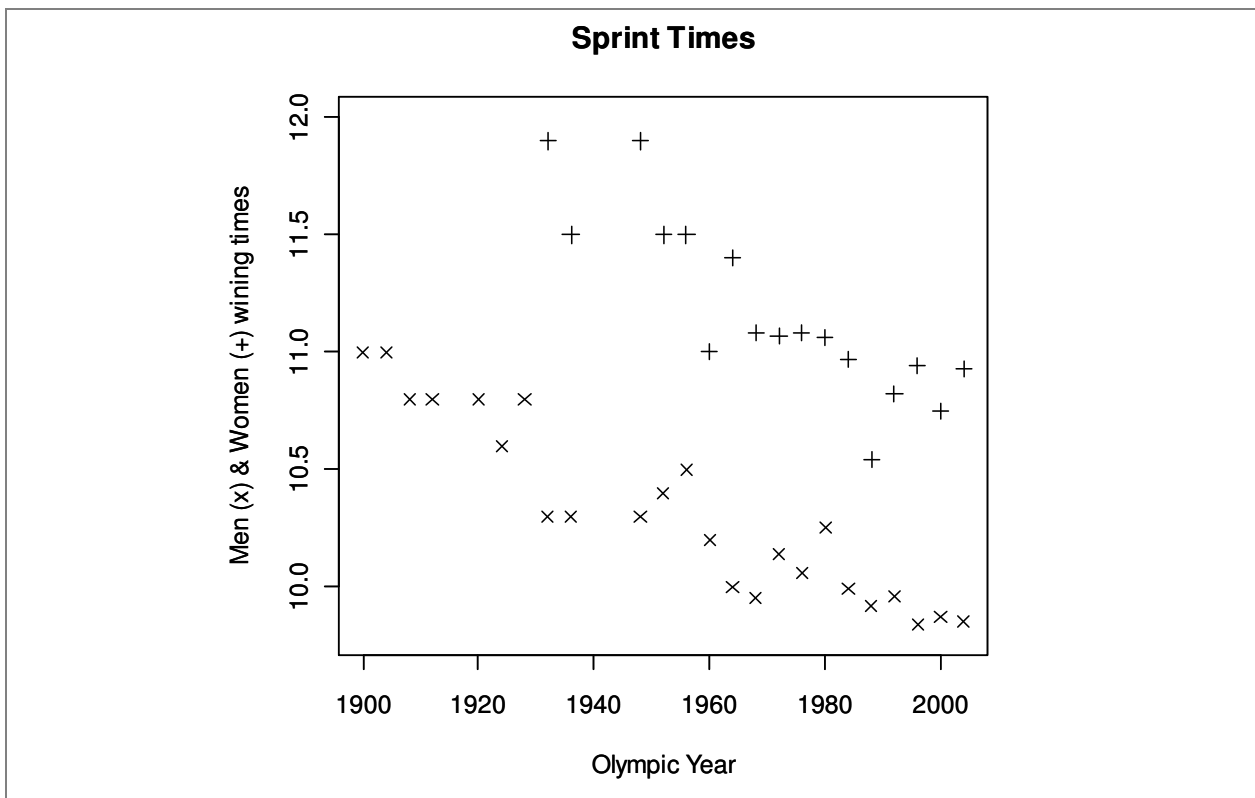
### סטטיסטיקה למדעי המחשב – פתרון תרגיל 3

1. הגדרת הנתונים ב-R:

```
sprinttimes=read.csv('sprinttimes.csv')
```

a. סרטוט תרשים הנתונים:

```
plot(sprinttimes$Olympic.year,sprinttimes$Men.s.winning.time..s., pch=4, ylab="Men (x) & Women (+) wining times", xlab="Olympic Year", main="Sprint Times", ylim=c(9.8,12))  
par(new=T)  
plot(sprinttimes$Olympic.year,sprinttimes$Women.s.winning.time..s., pch=3, ylab="Men (x) & Women (+) wining times", xlab="Olympic Year", main="Sprint Times", ylim=c(9.8,12))
```



ניתן לדמיין קשר ליניארי כלשהו בין השנה לזמני הריצה: אם נתעלם מ"רעש" מסוים בגרף נראה מגמה בה ככל שהשנים מתקדמות, זמני הריצה של המנצחים יורדים.

b. מציאת קווים חסינים:

```
> rline(sprinttimes$Olympic.year,sprinttimes$Men.s.winning.time..s.)  
[..]  
$slope  
[1] -0.01162162  
[..]  
> rline(sprinttimes$Olympic.year,sprinttimes$Women.s.winning.time..s.)  
[..]  
$slope  
[1] -0.01586538  
[..]
```

המשמעות של הנתונים היא שלאורך השנים, זמני הריצה של המנצחים נמצאים במגמת ירידה. כמו כן, מגמת הירידה בזמנים אצל הנשים תלולה יותר מאשר אצל הגברים (הנשים משפרות את הביצועים יותר מהגברים).

c. מציאת קווי הריבועים הפחותים:

```
ols.line=function(x,y){
  x<-x[!is.na(y)]#ignore data with missing y values
  y<-y[!is.na(y)]#ignore data with missing y values
  sxy=sum( (x-mean(x) ) * (y-mean(y) ) )
  sxx=sum( (x-mean(x))^2 )
  b1=sxy/sxx
  a1=mean(y)-b1*mean(x)
  return(list(slope=b1,intercept=a1))
}
> ols.line(sprinttimes$Olympic.year,sprinttimes$Men.s.winning.time..s.)
$slope
[1] -0.01100556
$intercept
[1] 31.82645
> ols.line(sprinttimes$Olympic.year,sprinttimes$Women.s.winning.time..s.)
$slope
[1] -0.01682207
$intercept
[1] 44.34705
```

בשיטת הריבועים הפחותים השיפוע של הנשים יותר תלול.

d. ע"פ מסקנות כותבי המאמר, בעוד כ-145 שנים, נשים ישיגו שיאי ריצה מהירים יותר מגברים. אינני מסכים עם טענה זו כיוון שכאשר מסתכלים על קו ליניארי המתאר נתונים מסוימים, לא ניתן להסיק ממנו אל מחוץ הנתונים שלנו. נוכל לראות דוגמה לכך אם נסתכל על פונקציה  $e^x$  בקטע  $[0,2]$ . בקטע זה נוכל לתאר קו ליניארי, אך כמובן שקו זה אינו מתאר את הנתונים אל מחוץ לתחום בו התבוננו. הסקה מחוץ לתחום במקרה זה תניב מסקנות שגויות.

e. **הקו החסין:**

$$33.04203 - 0.01162162x = 42.47135 - 0.01586538x \Rightarrow 0.00424376x = 9.42932 \Rightarrow x = 2221$$

ע"פ הקו החסין, המהפך יתבצע בשנת **2221**.

**קו הריבועים הפחותים:**

$$31.82645 - 0.01100556x = 44.34705 - 0.01682207x \Rightarrow 0.00581651x = 12.5206 \Rightarrow x = 2152$$

ע"פ קו הריבועים הפחותים, המהפך יתבצע בשנת **2152**.

2. טעינת הנתונים:

```
ex03data=read.csv('d:/stat/ex03data.data',header=T, sep=" ")
```

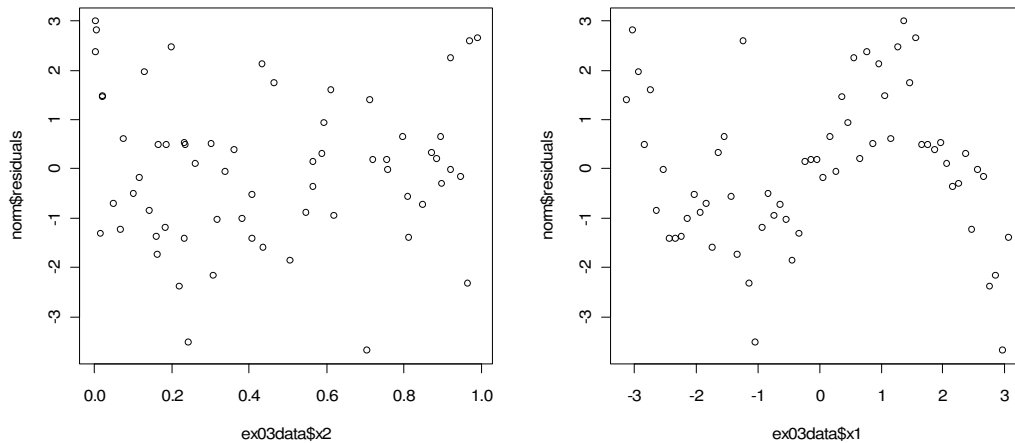
a. התאמת מישור ריבועים פחותים:

```
> lm(ex03data)
Call:
lm(formula = ex03data)

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

b. לכן, משוואת הניבוי היא  $y = 1.924 + 0.965 \cdot x_1 + 3.740 \cdot x_2$ .

c. נסתכל על השאריות של  $y$  ביחס ל- $x_1$  ו- $x_2$



ונציע את הפונקציה  $\sin$  עבור  $x_1$  ואת הפונקציה  $\cos$  עבור  $x_2$  (הפונקציה  $x^2$  גם תתאים ל- $x_2$ ).

d. נחשב את השונות המוסברת לפני הטרנספורמציות:

```
> orlm=lm(y~x1+x2,ex03data)
> orlm.coef=coef(orlm)
> orlm.pred=orlm.coef[1]+orlm.coef[2]*ex03data$x1+orlm.coef[3]*ex03data$x2
> r2_Original=r.squared(ex03data$y,orlm.pred)
> r2_Original
[1] 0.6760218
```

ולאחר הטרנספורמציות:

```
> f1= function (x1)
+ {
+ sin(x1)
+ }
> f2= function (x2)
+ {
+ cos(x2)
+ }
> trlm=lm(y~f1(x1)+f2(x2),ex03data)
> trlm.coef=coef(trlm)
> trlm.pred=trlm.coef[1]+trlm.coef[2]*f1(ex03data$x1)+trlm.coef[3]*f2(ex03data$x2)
> r2_trns=r.squared(ex03data$y,trlm.pred)
> r2_trns
[1] 0.8777983
```

הטרנספורמציות עזרו להסביר את השינוי במחירים כיוון שיצרו קשר טוב יותר בין מדד הזמן ומודעות למותג ובין המחיר. כך מתבטא גם באחוז השונות המוסבר שעלה מ-67.6% ל-87.77%.

3.

i. נוכל להניח שמקדם המתאם במקרים הבאים יהיה:

- קרוב ל-0 - הקשר בין הנתונים אינו ליניארי.
- קרוב ל-1 - הקשר בין הנתונים ליניארי וישר (לא הפוך).
- קרוב ל-0 - הקשר בין הנתונים אינו ליניארי.
- קרוב ל-0 - הקשר בין הנתונים אינו ליניארי.
- קרוב ל-0 - הקשר בין הנתונים אינו ליניארי.
- שלילי ומתקרב ל-(-1) - הקשר בין הנתונים ליניארי הפוך באופן יחסי אך אינו חזק כמו בדוגמה B למשל.

.ii

- a. הקשר אינו ליניארי, לכן, אחוז הנתונים המוסבר על ידי קשר זה יהיה **נמוך**.  
 b. הקשר בין הנתונים הוא ליניארי (אפילו ליניארי חזק) ולכן אחוז הנתונים המוסבר על ידי קשר זה יהיה **גבוה מאוד**.  
 c. גם כאן הקשר אינו ליניארי ולכן אחוז הנתונים המוסבר על ידי קשר זה יהיה **נמוך**.  
 d. הקשר בין הנתונים בגרף זה אינו ליניארי וייתכן שאינו יסביר אפילו אף נתון. לכן, אחוז הנתונים המוסבר על ידי קשר זה יהיה **נמוך מאוד** (ואף 0%).  
 e. בגרף זה גם לא קיים קשר ליניארי, אך בשל צורת הענן, אם נגדיר קשר כזה הוא צפוי להסביר חלק כלשהו מהנתונים (אשר יהיו במקרה קרובים אליו). לכן אחוז הנתונים המוסבר יהיה **נמוך עד בינוני** (ייתכן ויסביר עד כרבע מהנתונים).  
 f. בגרף זה קיים קשר ליניארי הפוך כלשהו. קשר זה אינו מובהק כמו בגרף B אך הוא קיים. לכן אחוז הנתונים המוסבר יהיה ככל הנראה **בינוני-גבוה** (מעל למחצית).

.iii

- a. הפעלת טרנספורמציה זו תקרב מאוד את מקדם המתאם ל-1 כיוון שהנתונים יוצרים צורה של פרבולה.  
 b. היות ואנו מפעילים את אותה הטרנספורמציה על נתוני X ו-Y ובשל הקשר הליניארי של הנתונים, הפעלת הטרנספורמציה שקולה לשינוי הסקאלה של הנתונים בלבד. כלומר, הקשר הליניארי בין הנתונים לא נפגע ולכן, מקדם המתאם אינו צפוי להשתנות במידה משמעותית.  
 c. בגרף G, ניתן לדמיין קשר ליניארי כלשהו בין הנתונים. הטרנספורמציה מבטאת קשר שאינו ליניארי (אלא קשר פולינומי). לכן הפעלת הטרנספורמציה צפויה להפחית את הקשר בין הנתונים ומכך שמקדם המתאם יתקרב ל-0.

.4

a. פונקצית ההסתברות המעריכית היא:

$$F_X(x) = \int_{-\infty}^x f_X(x) dx = \int_{-\infty}^0 f_X(x) dx + \int_0^x f_X(x) dx = \int_{-\infty}^0 0 \cdot dx + \int_0^x \lambda e^{-\lambda x} dx =$$

$$= 0 + -e^{-\lambda x} \Big|_0^x = -e^{-\lambda x} + e^0 = 1 - e^{-\lambda x}$$

b. תוחלת ההתפלגות המעריכית היא:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx = \int_{-\infty}^0 0 \cdot dx + \int_0^{\infty} \lambda x e^{-\lambda x} dx = \lambda \int_0^{\infty} x e^{-\lambda x} dx = \left[ \begin{array}{l} u(x) = x \mid v(x) = -\frac{e^{-\lambda x}}{\lambda} \\ u'(x) = 1 \mid v'(x) = e^{-\lambda x} \end{array} \right] =$$

$$= -x e^{-\lambda x} \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} = -x e^{-\lambda x} \Big|_0^{\infty} - \frac{e^{-\lambda x}}{\lambda} \Big|_0^{\infty} = 0 - 0 + 0 + \frac{e^0}{\lambda} = \frac{1}{\lambda}$$

c. נחשב את השונות המעריכית בשלבים:

$$E[X^2] = \int_{-\infty}^{\infty} x^2 \cdot f_X(x) dx = \int_{-\infty}^0 0 \cdot dx + \int_0^{\infty} \lambda x^2 e^{-\lambda x} dx = \int_0^{\infty} \lambda x^2 e^{-\lambda x} dx = \left[ \begin{array}{l} u(x) = x^2 \mid v(x) = -e^{-\lambda x} \\ u'(x) = 2x \mid v'(x) = \lambda e^{-\lambda x} \end{array} \right] =$$

$$= -x^2 e^{-\lambda x} \Big|_0^{\infty} + \frac{2}{\lambda} \int_0^{\infty} \lambda x e^{-\lambda x} dx = 0 + \frac{2}{\lambda} \cdot \frac{1}{\lambda} = \frac{2}{\lambda^2}$$

$$V(X) = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

d. הוכחת השוויון:

$$P(X \geq x+t \mid X > x) = \frac{P(X > x \mid X \geq x+t) P(X \geq x+t)}{P(X > x)} = \frac{P(X \geq x+t)}{P(X > x)} = \frac{1 - P(X < x+t)}{1 - P(X \leq x)} =$$

$$= \frac{1 - 1 + e^{-\lambda(x+t)}}{1 - 1 + e^{-\lambda x}} = \frac{e^{-\lambda(x+t)}}{e^{-\lambda x}} = e^{-\lambda t} = 1 - 1 + e^{-\lambda t} = 1 - [1 - e^{-\lambda t}] = 1 - P(X \leq t) = P(X \geq t)$$

a. הסיכוי שאורך הצלע קטן מ- $\alpha$  כלשהי הוא  $P(X \leq \alpha) = F_X(\alpha) = \alpha$  (בהנחה ש- $0 \leq \alpha \leq 1$ ). עבור  $\alpha < 0$  הסיכוי הוא 0 ועבור  $\alpha > 1$  הסיכוי הוא 1 (כמובן).

כפי שהוגדר בכיתה, פונקציית צפיפות ההסתברות עבור משתנה אחיד סטנדרטי היא:

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{אחרת} \end{cases} = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{אחרת} \end{cases}$$

b. על מנת ששטח הריבוע יהיה קטן מ- $\alpha$ , על אורך הצלע להיות קטן מ- $\sqrt{\alpha}$  ולכן:

$$P(Y \leq \alpha) = P(X \leq \sqrt{\alpha}) = F_X(\sqrt{\alpha}) = \begin{cases} \sqrt{\alpha} & 0 \leq x \leq 1 \\ 1 & x > 1 \\ 0 & x < 0 \end{cases}$$

פונקציית הצפיפות של Y היא:

$$f_Y(x) = F'_Y(x) = \begin{cases} 1 & x > 1 \\ \sqrt{\alpha} & 0 \leq x \leq 1 \\ 0 & x < 0 \end{cases}' = \begin{cases} \frac{1}{2\sqrt{x}} & 0 \leq x \leq 1 \\ 0 & \text{אחרת} \end{cases}$$

c. תוחלת הריבוע Y היא:

$$E[Y] = \int_{-\infty}^{\infty} y \cdot f_Y(y) dy = \int_{-\infty}^0 0 \cdot dy + \int_0^1 y \cdot \frac{1}{2\sqrt{y}} dy + \int_1^{\infty} 0 \cdot dy = 0 + \frac{1}{2} \int_0^1 \sqrt{y} dy =$$

$$= \frac{1}{3} y^{\frac{3}{2}} \Big|_0^1 = \frac{1}{3} 1^{\frac{3}{2}} - \frac{1}{3} 0^{\frac{3}{2}} = \frac{1}{3}$$

d. נחשב את שונות Y בשלבים:

$$E[Y^2] = \int_{-\infty}^{\infty} y^2 \cdot f_Y(y) dy = \int_{-\infty}^0 0 \cdot dy + \int_0^1 y^2 \cdot \frac{1}{2\sqrt{y}} dy + \int_1^{\infty} 0 \cdot dy = 0 + \frac{1}{2} \int_0^1 y^{\frac{3}{2}} dy =$$

$$= \frac{1}{5} y^{\frac{5}{2}} \Big|_0^1 = \frac{1}{5} 1^{\frac{5}{2}} - \frac{1}{5} 0^{\frac{5}{2}} = \frac{1}{5}$$

$$V(Y) = E[Y^2] - (E[Y])^2 = \frac{1}{5} - \frac{1}{9} = \frac{4}{45}$$

e. הסיכוי ש-Z קטן מ- $\alpha$  (עבור ערכים הגדולים מ-0) הוא:

$$P(Z \leq \alpha) = P(-\ln(X) \leq \alpha) = P(\ln(X) \geq -\alpha) = P(X \geq e^{-\alpha}) = 1 - P(X \leq e^{-\alpha}) = 1 - e^{-\alpha}$$

לכן, פונקציית הצפיפות של Z היא:

$$f_Z(x) = F'_Z(z) = \begin{cases} 1 - e^{-x} & 0 \leq z \\ 0 & z < 0 \end{cases}' = \begin{cases} e^{-x} & 0 \leq z \\ 0 & z < 0 \end{cases}$$

f. Z הוא התפלגות מערכית עם פרמטר 1. לכן, אם נציב פרמטר זה בנוסחאות שמצאנו בשאלה הקודמת, נראה שהתוחלת והשונות שלו, שניהם שווים ל-1.