בדיקת קורולציה בין משתנים סביבתיים על אורך החיים בישראל

סטטיסטיקה יישומית

מטרת המודל היא בדיקת ההשפעה של מספר גורמים: מספר תאונות הדרכים, אחוז בעלי תואר אקדמי מבני 25-54, מקרים חדשים של סרטן (כל הסוגים) , 2005-2009 (שיעור מתוקנן ל- 100,000 תושבים, שני המינים), זיהום כפי שנמדד על פי פליטת NOX - כלל תחמוצת החנקן (ממוצע שנתי - מק"ג/מ"ק) ודירוג אשכול חברתי – כלכלי על אורך תוחלת החיים בשנים בערים בישראל.

הצפי הוא שככל שיש יותר תאונות דרכים, זיהום ומקרים של סרטן, כך תוחלת החיים נמוכה יותר. לעומת זאת נצפה שככל שהדירוג במדד החברתי-כלכלי ואחוז שיעור בעלי התואר האקדמי גובה יותר, כך תוחלת החיים ארוכה יותר.

לשם כך נלקח מדגם בכל אחת מן הקטגוריות ב-30 ערים אקראיות ברחבי הארץ. נתוני המדגם, התקבלו מהלשכה המרכזית לסטטיסטיקה[1,2] והמשרד לאיכות הסביבה[3]*.

---

[1] יישובים וחלוקות גאוגרפיות אחרות, הלשכה המרכזית לסטטיסטיקה.
https://www.cbs.gov.il/he/settlements/Pages/default.aspx?subject=תמותה ותוחלת חיים

[2] אפיון יחידות גאוגרפיות וסיווגן לפי הרמה החברתית-כלכלית של האוכלוסייה בשנת 2015
https://www.cbs.gov.il/he/publications/Pages/2019/אפיון-יחידות-גאוגרפיות-וסיווגן-לפי-הרמה-החברתית-כלכלית-של-האוכלוסייה-בשנת-2015.aspx

[3] דוח שנתי תחנות כלליות, המשרד להגנת הסביבה.
https://www.sviva.gov.il/infoservices/reservoirinfo/doclib2/publications/p0801-p0900/p0837-tables.pdf

*יש לציין כי במידה לעיר מסויימת לא היה נתונים, נלקחו הנתונים מהאזור הגאוגרפי הקרוב ביותר.

המודל והנחותיו:

Y – גורם מוסבר – אורך תוחלת החיים ב-30 ערים במדינת ישראל.

גורמים מסבירים:

X_1 - מספר תאונות הדרכים בעיר מסויימת.

X_2 - אחוז בעלי תואר אקדמי מבני 25-54 בעיר מסויימת מתוך סך כל תושבי העיר.

X_3 - מקרים חדשים של סרטן (כל הסוגים) (שיעור מתוקנן ל- 100,000 תושבים), שני המינים.

X_4 - זיהום כפי שנמדד על פי פליטת NOX - כלל תחמוצת החנקן (ממוצע שנתי - מק"ג/מ"ק).

X_5 – דירוג העיר על פי המדד האשכול חברתי – כלכלי.

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_3 + \widehat{\beta_4}X_4 + \widehat{\beta_5}X_5 + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$n = 30$$

בדיקת תלות בין המשתנים המסבירים:

$$(x^T x)^{-1} =$$

```
>matrix1<-as.matrix(data.frame(x0,x1,x2,x3,x4,x5))
>matrix2<-solve(t(matrix1)%*%matrix1)
```

מטריצה לבדיקת מולטיקולונריות						
	x_0	x_1	x_2	x_3	x_4	x_5
x_0	3.19E+00	-2.59E-05	-5.68E-01	-9.47E-03	-1.05E-02	1.75E-01
x_1	-2.59E-05	1.84E-08	5.38E-05	5.80E-08	-9.71E-08	-3.30E-06
x_2	-5.68E-01	5.38E-05	5.41E+00	1.76E-03	-2.70E-03	-3.09E-01
x_3	-9.47E-03	5.80E-08	1.76E-03	3.08E-05	2.53E-05	-6.61E-04
x_4	-1.05E-02	-9.71E-08	-2.70E-03	2.53E-05	2.38E-04	-1.01E-03
x_5	1.75E-01	-3.30E-06	-3.09E-01	-6.61E-04	-1.01E-03	3.71E-02

ניתן לראות כי באלכסון המטריצה אף ערך איננו גדול מ-10 ועל כן אין תלות בין הגורמים המסבירים.

$\underline{\beta} = (x^T x)^{-1} x^T \underline{y} =$

```
>Bmatrix<-matrix2%*%matrix3
```

	$\underline{\beta}$
β_0	*7.848245e+01*
β_1	*-4.190225e-05*
β_2	*1.124303e+01*
β_3	*-3.793435e-04*
β_4	*3.404147e-04*
β_5	*-1.445402e-01*

משמעות הבטאות שהתקבלו:

β_0 – מנרמל את הערכים (הרי אין זה הגיוני שאם שיעור בעלי התואר האקדמאי הוא 0 אז אורך תוחלת החיים היא 0 שנים), גם אם כל המשתנים אפסיים עדיין תוחלת החיים המשוערת היא $\mathbf{7.848245}e + \mathbf{01}$ שנים.

β_1 – על כל עלייה אחת בכמות תאונת הדרכים בעיר מסויימת יורדת תוחלת החיים הממוצעת באותה עיר בשיעור של $\mathbf{4.190225e - 05}$ שנים.

β_2 - על כל עלייה באחוז אחד בשיעור בעלי תואר אקדמאי מסך התושבים בעיר מסויימת, כך ישנה עלייה של כ-$\mathbf{1.124303e + 01}$ שנים באורך תוחלת החיים הממוצע באותה עיר.

β_3 – על כל עלייה במקרה סרטן אחד פר 100,000 תושבים ישנה ירידה של $\mathbf{3.793435e - 04}$ שנים באורך תוחלת החיים הממוצע באותה עיר.

β_4 – על כל עלייה של יחידה אחת בריכוז תחמוצת החנקן בממוצע לשנה לעיר, ישנה ירידה של $\mathbf{3.404147}e - \mathbf{04}$ שנים באורך תוחלת החיים הממוצע באותה עיר.

β_5 – על כל עלייה של יחידה אחת בדירוג האשכול החברתי-כלכלי של אותה עיר ישנה ירידה של $\mathbf{1.445402e - 01}$ שנים באורך תוחלת החיים הממוצע באותה עיר. ניתן לראות שזה נוגד את הצפייה שדווקא עלייה במדד תוביל לעלייה בתוחלת החיים ואילו על פי המודל ישנה דווקא ירידה.

כלומר המודל שהתקבל הוא:

$$\hat{Y} = (7.848245e + 01) + (-4.190225e - 05)X_1 + (1.124303e + 01)X_2 + (-3.793435e - 04)X_3 + (3.404147e - 04)X_4 + (-1.445402e - 01)X_5$$

<u>בדיקת השערות:</u>

$\mathbf{H_0}: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

$\mathbf{H_1}: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 \neq 0$

```
> anovatable<-aov(y~matrix1.1)
> summary.aov(anovatable)
            Df Sum Sq Mean Sq F value  Pr(>F)
matrix1.1    5  55.16   11.03   5.381 0.00186 **
Residuals   24  49.20    2.05
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P<0.05, מכאן שנדחה את H_0 ברמת מובהקות 5% ונסיק כי לפחות אחד מהגורמים המסבירים הוא כן מובהק.

נבדוק איזה מהבטאות אינן מובהקות:

```
> lm1<-lm(y~matrix1.1)
> summary(lm1)

Call:
lm(formula = y ~ matrix1.1)

Ressiduals:
    Min      1Q  Median      3Q     Max 
-2.5615 -0.8598  0.1172  0.3908  2.6036 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 78.4824494  2.5577397  30.684   <2e-16 ***
matrix1.1x1 -0.0000419  0.0001941  -0.216   0.8309    
matrix1.1x2 11.2430296  3.3293870   3.377   0.0025 ** 
matrix1.1x3 -0.0003793  0.0079520  -0.048   0.9623    
matrix1.1x4  0.0003404  0.0220818   0.015   0.9878    
matrix1.1x5 -0.1445402  0.2759543  -0.524   0.6052    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.432 on 24 degrees of freedom
Multiple R-squared:  0.5285,	Adjusted R-squared:  0.4303 
F-statistic: 5.381 on 5 and 24 DF,  p-value: 0.001859
```

ניתן לראות כי ישנם שתי בטאות שיצאו מובהקות עם Pvalue<0.05:

β_0 : Pvalue = <2e-16

β_2 : Pvalue = 0.0025

כלומר β_0 – שמנרמל את הערכים (הרי אין זה הגיוני שאם שיעור בעלי התואר האקדמאי הוא 0 אז אורך תוחלת החיים היא 0 שנים), ו- β_2 - אחוז בעלי תואר אקדמי מבני 25-54 מתוך סך תושבי העיר הם מובהקים (בעלי משמעות). דהיינו אחוז בעלי תואר אקדמי מבני 25-54 מתוך סך תושבי העיר משפיע על אורך תוחלת החיים באותה עיר.

לעומת זאת β_1 - מספר תאונות הדרכים, β_3 - מקרים חדשים של סרטן (כל הסוגים) (שיעור מתוקנן ל- 100,000 תושבים) שני המינים, β_4 - זיהום כפי שנמדד על פי פליטת NOX - כלל תחמוצת החנקן (ממוצע שנתי - מק"ג/מ"ק) ו- β_5 - דירוג על פי המדד האשכול חברתי – כלכלי, אינם משפיעים באופן מובהק על אורך תוחלת החיים.

בנוסף ניתן לראות כי קיבלנו:

$R^2{}_{adj} = 0.4303$

על כן ניתן להסיק כי קיים קשר לינארי בינוני בין אורך תוחלת החיים בעיר מסויימת בישראל לשיעור בעלי התואר האקדמי בקרב בני 25-52 מתוך סך תושבי אותה עיר.

לכן נשאר עם המודל הבא:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_2}\mathrm{X}_2 =$$

$$\hat{Y} = (7.848245e + 01) + (1.124303\mathrm{e} + 01)\mathrm{X}_2$$

מדובר במודל רגרסיה פשוטה, כאשר Y הוא הגורם המוסבר שהוא אורך תוחלת החיים בעיר מסויימת בישראל ו-X_2 הוא הגורם המסביר שהוא שיעור בעלי התואר האקדמי בקרב בני 54-25 מתוך סך תושבי אותה עיר.

<u>משמעות הבטאות שנשארו:</u>

β_0 – מנרמלת את הערכים כך שגם שיעור בעלי התואר האקדמי בקרב בני 54-25 הוא 0, תוחלת החיים הצפויה בעיר כל שהיא בישראל היא $\mathbf{7.848245}\boldsymbol{e} + \mathbf{01}$ שנים.

β_2 - על כל עלייה באחוז אחד בשיעור בעלי תואר אקדמאי בקרב בני 54-25 מתוך סך התושבים בעיר מסויימת, כך ישנה עלייה של כ-$\mathbf{1.124303e} + \mathbf{01}$ שנים באורך תוחלת החיים הממוצע באותה עיר.

חיזוי:

ניקח מספר שרירותי של 10% בעלי תואר אקדמאי בגילאים 25-54 בעיר כל שהיא בישראל.

$X_0 = 0.1$

$\hat{Y} = (7.848245e + 01) + (1.124303e + 01) * 0.1 = 79.606753$

נבצע רווח סמך ל$E(y)$:

$$\hat{Y} \pm t_{n-k-1,1-\frac{\alpha}{2}} \sqrt{(\frac{1}{n} + \frac{(\bar{X} - X_0)^2}{S_{xx}})MSE}$$

```
> #MSE חישוב
> SSE<-sum((as.matrix(y)-as.matrix(Y_model))^2)
> MSE<-SSE/(30-2)
[1] 2.828268
> #Sxx חישוב
> sd(x2)
[1] 0.144982
```

MSE = 2.828268
S_{xx}=0.144982
$t_{30-1-1,0.975}$ = 2.048

$$79.606753 \pm 2.048\sqrt{\left(\frac{1}{30} + \frac{(0.30402-0.10)^2}{0.144982}\right)2.639717} = 79.606753 \pm$$

$$1.883547937 = [77.72320506, 81.49030094]$$

כלומר אדם המתגורר בעיר ובה 10% מתושבי העיר הם בעלי תואר אקדמי בגילאים 25-54 יחיה בממוצע בין 77.7232050 ל-81.49030094 שנים.

נבצע חיזוי נקודתי לעיר כל שהיא ובה שיעור בעלי התואר האקדמי בני 25-54 הוא כ-12%:

$$\hat{Y} \pm t_{n-k-1,1-\frac{\alpha}{2}} \sqrt{(1 + \frac{1}{n} + \frac{(\bar{X} - X_0)^2}{S_{xx}})MSE}$$

X_0 = 0.12%

$\hat{Y} = (7.848245e + 01) + (1.124303e + 01) * 0.12 = 79.8316136$

$$79.8316136 \pm 2.048\sqrt{\left(1 + \frac{1}{30} + \frac{(0.30402-0.12)^2}{0.144982}\right)2.639717} = 79.8316136 \pm$$

$$3.745245367 = [76.08636823, 83.576858967]$$

כלומר אורך החיים הצפוי לאדם המתגורר בעיר ובה 12% מסך תושבי העיר הם בעלי תואר אקדמי בגילאים 25-54 הוא בין 76.08636823 ל-83.576858967 שנים.

ניתוח שהטעות מתפלגת נורמלית:

$H_0: e \sim N(\mu = 0, \sigma^2 = MSE = 2.828268)$

H_1: טעות לא מתפלגת נורמלית

```
> ei=as.matrix(y)-as.matrix(Y_model)
> ks.test(ei,"pnorm", 0, MSE)

        One-sample Kolmogorov-Smirnov test

data:  ei
D = 0.34159, p-value = 0.001254
alternative hypothesis: two-sided
```

P<0.05, לכן נדחה H_0 ברמת מובהקות 5% ונסיק כי הטעות אינה מתפלגת נורמלית עם תוחלת של **0** ושונות של **2.828268**.

מסקנות המחקר:

מצאנו כי מכל הגורמים שהנחנו בעלי השפעה על אורך תוחלת החיים בישראל רק שיעור בעלי התואר האקדמי מתוך סך תושבי העיר הוא גורם בעל קשר לינארי לאורך תוחלת החיים בישראל ועוצמת הקשר היא בינונית.

אם זאת, מכיוון שהטעות אינה מתפלגת נורמלית יש להניח כי השתמשנו בכלי ניתוח לא מתאימים ועל כן על מנת לבדוק בצורה מיטבית יותר את הקשר הלינארי יש לחזור על המחקר בעזרת שיטות ניתוח סטטיסטיות מתאימות יותר.