

# כלים כמותיים מתקדמים של תכן סטטיסטי לאיכות

פרק 3 - רגרסיה מרובה

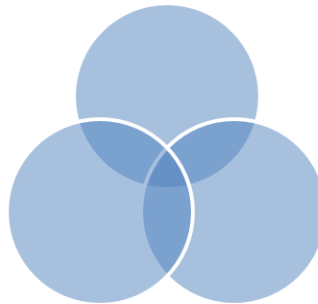
תוכן העניינים

1. כללי ..... 1

## רגרסיה מרובה:

### רקע:

ניבוי המשתנה התלוי באמצעות יותר ממשתנה ב"ת אחד.  
 המודל באוכלוסייה:  $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$ .  
 מקדמי מודל הרגרסיה המרובה:  
 $\alpha$  = חותך אחד שמשמעותו: הציון המנובא כאשר כל המשתנים הב"ת = 0.  
 $\beta_1 \dots \beta_j$  = מקדמי השיפוע. מס' הבטות = למספר המשתנים הב"ת במודל.  
 $\beta_j$ : ההשפעה הייחודית של המשתנה הב"ת המסוים לניבוי  
 המשתנה התלוי, בניכוי השפעתם של כל יתר המשתנים הב"ת האחרים המצויים  
 במשוואת הרגרסיה.



### אמידת מודל הרגרסיה המרובה:

ברגרסיה מרובה, כמו ברגרסיה פשוטה, שיטת האמידה הטובה ביותר היא שיטת הריבועים הפחותים. כלומר, נרצה להביא את סכום הטעויות בניבוי למינימום.  
 מפתרון פונקציית הריבועים הפחותים נקבל את אומדי הרגרסיה:  $\hat{\alpha}, \hat{\beta}_1 \dots \hat{\beta}_j$ .

### מבחני מובהקות:

1. מבחן F למובהקות הרגרסיה:  
 בדיקה האם קיים קשר ליניארי בין המשתנה התלוי Y לבין לפחות אחד מהמשתנים המסבירים.

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

ההשערות הן:  $H_1 : \text{Not } H_0 = \text{at least one of the } \beta\text{'s is not 0}$

טבלת ניתוח שונות (ANOVA):

| מקור           | סכום ריבועים<br>SS | דרגות חופש<br>$d.f.$ | ממוצע סכום ריבועים<br>$MS = \frac{SS}{d.f.}$ | $F_{st} \sim F_{k,n-k-1}$  |
|----------------|--------------------|----------------------|----------------------------------------------|----------------------------|
| מודל<br>הגרסיה | SSR                | K                    | $MSR = \frac{SSR}{K}$                        | $F_{st} = \frac{MSR}{MSE}$ |
| שאריות         | SSE                | $n-k-1$              | $MSE = \frac{SSE}{(n-k-1)}$                  |                            |
| סה"כ           | TSS                | $n-1$                |                                              |                            |

סטטיסטי המבחן:  $F_{st} = \frac{MSR}{MSE}$

כלל הכרעה: נדחה את  $H_0$  אם:  $F_{st} \geq F_{k,n-k-1}^{1-\alpha}$

חישוב סכומי הריבועים:

$$TSS = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

$$SSR = R^2 \cdot TSS$$

$$SSE = (1 - R^2)TSS$$

פרופורציית השונות המוסברת  $R^2$ :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

בגרסיה מרובה אומד זה לפרופורציית השונות המוסברת הוא בעייתי שכן הוא מושפע ממספר המשתנים הב"ת במודל. אומד זה יכול רק לגדול בהוספת משתנים ב"ת למודל ולכן לא ייתן לנו אינדיקציה האם כדאי היה לי להוסיף אותם למודל או לא.

האומד המתוקן לפרופורציית השונות המוסברת  $AdjR^2$ :

$$\bar{R}^2 = 1 - \left[ \frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

בניגוד ל- $R^2$  לוקח בחשבון את מספר המשתנים הב"ת במודל. יכול שלא לגדול ואף לקטון בהוספת משתנה ב"ת שלא תורם תרומה משמעותית לניבוי.

2. מבחן  $t$  למובהקות משתנה ב"ת יחיד :

$$\begin{aligned} H_0 : \beta_j &= 0 \\ H_1 : &else \end{aligned}$$

השערות :

סטטיסטי המבחן וכלל הכרעת השערת האפס :

$$|t_{\hat{\beta}_j}| = \left| \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}} \right| > t_{\left(T-k-1; 1-\frac{\alpha}{2}\right)}$$

$$\hat{\beta}_j \pm t_{n-k-1, 1-\frac{\alpha}{2}} \cdot s.e.(\hat{\beta}_j) : \beta_j \text{ לאמידת ה-}$$

### שאלות:

- 1) לצורך בדיקת ההשערה שקיים קשר בין מספר המוניות בעיר באר שבע ( $y$ ) לבין מספר התושבים בעיר באלפים ( $x_1$ ) ומספר הרכבים הפרטיים באלפים ( $x_2$ ). הוחלט לבנות מודל רגרסיה מהצורה:  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ , על סמך הנתונים הבאים:  $\sum y_i = 1673$ ,  $\sum y_i^2 = 338,657$ ,  $MSE = 119.789$ .
- א. ע"ס הנתונים הנ"ל, השלימו את טבלת ניתוח השונות הבאה.
- איזו השערה ניתן לבדוק באמצעותה? כתוב את ההשערה ובחן אותה.

| SOURCE     | SS | DF | MS | F |
|------------|----|----|----|---|
| Regression |    |    |    |   |
| Error      |    |    |    |   |
| Total      |    | 8  |    |   |

- ב. חשבו את מדד טיב ההתאמה. הסבר את משמעותו.
- ג. נתונה טבלת המקדמים (החלקית) הבאה :

|           | Coefficients | Standard Error | t Stat   | P-value | Lower 95% | Upper 95% |
|-----------|--------------|----------------|----------|---------|-----------|-----------|
| Intercept | -511.727     | 114.9476       |          |         |           |           |
| X 1       | 9.208785     |                | 3.732167 |         |           |           |
| X 2       | -8.79921     | 4.420456       |          |         |           |           |

- i. רשמו את האומדן למשוואת הרגרסיה ופרשו את מקדמיה.
- ii. בחנו את ההשערה כי קיים קשר בין מספר הרכבים הפרטיים לבין מספר המוניות ברמת מובהקות של 5%.
- iii. בנו רווח סמך למקדם של מספר התושבים בעיר

- iv. ענה ללא חישוב (על סמך הסעיפים הקודמים) – האם קיים קשר בין מספר התושבים לבין מספר המוניות ברמת מובהקות 5%?
- v. מהי תחזית מס' המוניות בבאר שבע עבור 100,000 תושבים ו-52,000 מכוניות פרטיות?
- vi. האם ניתן לסמוך על תחזית זאת?

### תרגול מסכם:

- (2) מעוניינים למצוא קשר בין מחיר הדירה (ב-\$) לבין ארבעה משתנים מסבירים: (1) שטח הדירה ו-(2) גודל שטח האמבטיה (ב-Sqft) וכן (3) מרחק הדירה מהים ו-(4) מהאוניברסיטה (במיילים). לשם כך נדגמו מספר דירות והריצו רגרסיה אשר בה המשתנה המוסבר הוא מחיר הדירה. להלן פלט הרגרסיה שהתקבל:

Model Summary

| Model | R                 | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1     | .952 <sup>a</sup> |          |                   |                            |

a. Predictors: (Constant), Sea\_Dist, Apartment, Bath

ANOVA<sup>b</sup>

| Model |            | Sum of Squares | df | Mean Square | F | Sig.              |
|-------|------------|----------------|----|-------------|---|-------------------|
| 1     | Regression |                |    |             |   | .000 <sup>a</sup> |
|       | Residual   |                |    |             |   |                   |
|       | Total      | 1940484.615    | 25 |             |   |                   |

a. Predictors: (Constant), Univ\_Dist, Bath, Sea\_Dist, Apartment

b. Dependent Variable: Price

Coefficients<sup>a</sup>

| Model |            | Unstandardized Coefficients |            | Standardized Coefficients | t      | Sig. |
|-------|------------|-----------------------------|------------|---------------------------|--------|------|
|       |            | B                           | Std. Error | Beta                      |        |      |
| 1     | (Constant) | -265.514                    | 146.673    |                           | -1.810 | .085 |
|       | Apartment  |                             | .449       | .722                      | 6.572  |      |
|       | Bath       | 4.256                       |            | .297                      | 2.687  | .014 |
|       | Sea_Dist   | -32.114                     | 11.090     | -.223                     |        | .009 |
|       | Univ_Dist  | 11.746                      | 9.439      | .095                      | 1.244  | .227 |

a. Dependent Variable: Price

ענו על הסעיפים הבאים :

- א. מלאו את התאים החסרים בטבלה (אם לא ניתן למלא את כל התאים החסרים באופן מלא נמקו באופן מפורש מדוע לא ניתן).
- ב. כתבו את האומדן למשוואת מחיר הדירה בצורה מפורשת על סמך הפלט הנ"ל. פרשו את מקדמי הרגרסיה.
- ג. בדקו האם ארבעת הגורמים ביחד אכן מסבירים את מחיר הדירה. הסבירו את המסקנה שהגעתם אליה. השתמשו ברמת מובהקות 5%.
- ד. הסבירו מהו ערך ה-Pvalue ומה ניתן להסיק ממנו לגבי המשתנים המסבירים?
- ה. בנו רווח סמך למקדם גודל שטח האמבטיה. השתמשו ברמת מובהקות של 2%.
- ו. ברמת מובהקות של 5% יש לבדוק האם המרחק מהאוניברסיטה אכן משפיע על מחיר הדירה.
- ז. האם במודל הרגרסיה הנוכחי ניתן לוותר על גורם המרחק מהים? השתמשו ברמת מובהקות 1%.
- ח. בדקו את ההשערה כי קיים קשר חיובי בין גודל הדירה למחירה ברמת מובהקות של 5%.

### תשובות סופיות:

- 1) א. השערה:  $H_0 : \beta_1 = \beta_2 = 0$   
 $H_1 : \text{at least one of the } \beta\text{'s is not } 0$

| SOURCE     | SS        | DF | MS        | F       |
|------------|-----------|----|-----------|---------|
| Regression | 26945.784 | 2  | 13472.892 | 112.414 |
| Error      | 718.733   | 6  | 119.789   |         |
| Total      | 27664.517 | 8  |           |         |

- ב. 97.4% .i.  $\hat{y}_i = -511.727 + 9.208x_{1i} - 8.799x_{2i}$
- ii. אין עדות לכך. .iii  $p(3.17 \leq \beta_1 \leq 15.24) = 0.95$
- iv. כן. .v 0.321 .vi. כן.
- 2) א. ראו סרטון. ב.  $\hat{y}_i = -256.514 + 2.95x_{1i} + 4.256x_{2i} - 32.114x_{3i} + 11.746x_{4i}$
- ג. לפחות אחד מהמשתנים הב"ת שונה מאפס באוכלוסייה.
- ד. ראו סרטון. ה.  $p(1.016 \leq \beta_2 \leq 7.496) = 0.98$
- ו. לא. ז. לא. ח. יש עדות לכך.