ניתוח נתונים

פרק 28 - רגרסיה מרובה

תוכן העניינים

1. כללי 1

רגרסיה מרובה:

רקע:

ניבוי המשתנה התלוי באמצעות יותר ממשתנה ב"ת אחד.

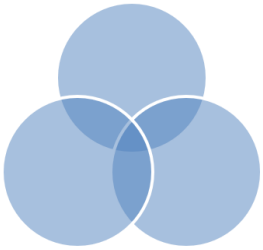
המודל באוכלוסייה: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$.

מקדמי מודל הרגרסיה המרובה:

α = חותך אחד שמשמעותו: הציון המנובא כאשר כל המשתנים הב"ת = 0.

$\beta_1 ... \beta_J$ = מקדמי השיפוע. מס' הבטות = למספר המשתנים הב"ת במודל.

משמעות מקדם השיפוע β_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: $\hat{\alpha}, \hat{\beta}_1 ... \hat{\beta}_J$.

מבחני מובהקות:

1. מבחן F למובהקות הרגרסיה:

בדיקה האם קיים קשר ליניארי בין המשתנה התלוי Y לבין לפחות אחד מהמשתנים המסבירים.

ההשערות הן:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : Not \quad H_0 : at\ least\ one\ of\ the\ \beta's\ is\ not\ 0$$

טבלת ניתוח שונות (ANOVA):

מקור	סכום ריבועים SS	דרגות חופש $d.f.$	ממוצע סכום ריבועים $MS=\frac{SS}{d.f.}$	$F_{st}\sim F_{k,n-k-1}$
מודל הרגרסיה	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:

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

ברגרסיה 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 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:

השערות: $H_0: \beta_j = 0$, $H_1: else$.

סטטיסטי המבחן וכלל הכרעת השערת האפס:

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

רווח בר סמך לאמידת ה- β_j: $\hat{\beta}_j \pm t_{n-k-1,1-\frac{\alpha}{2}} s.e.\left(\hat{\beta}_j\right)$.

3. מבחן F חלקי (partial F):

בודק את ההשערה שתוספת של משתנה אחד או קבוצה של משתנים מוסיפה תוספת מובהקת לניבוי המשתנה התלוי מעבר למשתנים אחרים שקיימים כבר במודל.

השערות: $H_0: \beta_1 = \beta_2 = ... = \beta_p = 0$, $H_1: otherwise$. $p < k$.

ביצוע המבחן:

מריצים שתי רגרסיות:

1. UR: המודל המלא – רגרסיה עם כל המשתנים הב"ת (K).
2. R: המודל החלקי – רגרסיה תחת H_0 ($K-P$).

סטטיסטי המבחן:

$$F = \frac{\frac{R_{UR}^2 - \frac{R_R^2}{p}}{\frac{1-R_{UR}^2}{n-k-1}} = \frac{SSR_{UR} - \frac{SSR_R}{p}}{\frac{1-SSR_{UR}}{n-k-1}} = \frac{SSE_R - \frac{SSE_{UR}}{p}}{\frac{1-SSE_{UR}}{n-k-1}}}{}.$$

כלל הכרעה: $F > f_{1-\alpha}^{p,n-k-1}$.

קשר בין מבחן F חלקי למבחן t:

קיים קשר בין מבחן F חלקי למובהקות תוספת משתנה ב"ת יחיד, למבחן t למובהקות אותו משתנה:

$$\begin{matrix} F > f_{1-\alpha}^{1,n-k-1} = t^2_{1-\frac{\alpha}{2},n-k-1} \\ pvalue = pvalue \end{matrix}.$$

מולטיקוליניאריות:

בעיית המולטיקוליניאריות:
מצב שבו המשתנים המסבירים הם בעלי מתאם גבוה בינם ולבין עצמם.
מתאמים גבוהים אלו יוצרים קושי ביכולת להבחין בהשפעה הבודדת של כל משתנה מסביר על המשתנה המוסבר.

כיצד מזהים בעיית מוליטיקוליניאריות?

- R^2 גדול וערכי t קטנים, או שדוחים את השערת האפס במבחן F אך לא במבחני t.
- מקדמי מתאם זוגיים גבוהים בין המשתנים המסבירים (לפי טבלת קורלציות).

איתור המולטיקולינאריות:

נבחר לרוב להוציא את אחד מהמשתנים לפי הקריטריונים הבאים:

1. המשתנה המתואם ביותר גם עם יתר המשתנים המסבירים.
2. המשתנה הכי פחות מובהק שהתקבל בהרצה הראשונית (לפי t).

אבחון קוליניאריות – מדד VIF (Variance Inflationary Factor):

מדד זה מבטא את הגידול (האינפלציה) בשונות של מקדם הרגרסיה של משתנה ב"ת מסוים כתוצאה מהמתאם שלו עם משתנים ב"ת אחרים המצויים במשוואה:

$$X_j = \frac{1}{\left(1-R_j^2\right)}.$$

R_j^2 הוא מקדם ההסבר של X_j, כל יתר המשתנים ב"ת.

כאשר: $VIF_j > 5$ המשמעות היא ש- X_j מתואם בצורה חזקה עם יתר המשתנים הב"ת במשוואה.

חישוב מובהקות התוספת (F חלקי) של משתנה ב"ת מסוים על פני האחרים:

במקרה של מולטיקוליניאריות במודל (מתאם חזק בין משתנים ב"ת) בכדי לדעת איזה משתנה ב"ת יש להוציא, ניתן לבחון את התוספת לניבוי של המשתנה ה"חשוד" על פני האחרים. אם היא איננה מובהקת, זוהי אינדיקציה שיש להוציאו מהמודל.

שאלות:

המודל ומבחני המובהקות:

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: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_{2i} + \varepsilon_i$, על סמך הנתונים הבאים: $\sum_i 1673$, $\sum_i y_i^2 = 338657$, $MSE = 119.789$.

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?

vii. חשב את סטטיסטי F חלקי של מס' הרכבים הפרטיים.
האם מובהק? (ענה ללא חישוב).

תרגול מסכם:

2) מעוניינים למצוא קשר בין מחיר הדירה (ב-$) לבין ארבעה משתנים מסבירים:

1. שטח הדירה.
2. גודל שטח האמבטיה (ב-Sqft).
3. מרחק הדירה מהים.
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[a]			

a. Predictors: (Constant), Sea_Dist, Apartment, Bath

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression					.000[a]
	Residual					
	Total	1940484.615	25			

a. Predictors: (Constant), Univ_Dist, Bath, Sea_Dist, Apartment

b. Dependent Variable: Price

Coefficients[a]

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

ענה על הסעיפים הבאים:

א. מלאו את התאים החסרים בטבלה (אם לא ניתן למלא את כל התאים החסרים באופן מלא ונמקו 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%.

ט. נתונה מטריצת מקדמי המתאם הבאה:

	X1	*X2*	*X3*	*X4*
X1	**1**			
X2	**0.228579**	**1**		
X3	**-0.22413**	**-0.13924**	**1**	
X4	**-0.24545**	**-0.97295**	**0.029537**	**1**

מה ניתן ללמוד ממנה ומה משמעותה לגבי המודל?

י. הניחו כי השאריות המתקבלות מניתוח הרגרסיה הן בעלות הערכים הבאים (סדר הקריאה הוא משמאל לימין) הנח כי אלו כל השאריות הקיימות במודל: $-1, -12, 3, 6, -4, 5, 10, -5, 6, 18, 12, -3, -7, -3, 9, 7$.

האם משתנים X_2 ו- X_4 מוסיפים תוספת משמעותית לניבוי? אם לא ניתן לענות על השאלה, ציין מדוע.

יא. מה יהיו תוצאות מבחן F לבדיקת התוספת לניבוי של המרחק מהאוניברסיטה על פני המשתנים האחרים (ענה ללא חישוב).

תשובות סופיות:

1) א.

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

$H_0: \beta_1 = \beta_2 = 0$
$H_1: at\ least\ one\ of\ the\ \beta's\ is\ not\ 0$, נדחה את השערת האפס.

ב. 97.4% ג.i. $\hat{y}_i = -511.727 + 9.208x_{1i} - 8.799x_{2i}$.

ii. אין עדות לכך. iii. $p(3.17 \le \beta_1 \le 15.24) = 0.95$.

iv. מובהק. v. 321 מוניות. vi. כן. vii. $F = 3.96$.

2) א.1. 0.907 א.2. 0.89 א.3. 92 א.4. 1760019.5

א.5. 180465 א.6. 4 א.7. 21 א.8. 440004.875

א.9. 8593.57 א.10. 51.2 א.11. 2.95 א.12. $\text{sig} < 0.001$.

א.13. 1.58 א.14. -2.896.

ב. $\hat{y}_i = -256.514 + 2.95x_{1i} + 4.256x_{2i} - 32.114x_{3i} + 11.746x_{4i}$.

ג. ראו סרטון. ד. ראו סרטון. ה. $p(1.016 \le \beta_2 \le 7.496) = 0.98$.

ו. אין עדות לכך. ז. לא. ח. יש עדות לכך.

ט. בעיית קוליניאריות. י. ראו סרטון. יא. ראו סרטון.