תאריך המבחן:26.8.2014
שם המרצה: פרופ' הלל בר גרא
שם הקורס: מודלים של רגרסיה לינארית
מספר הקורס:364-1-1061
שנה: סמסטר ב' תשע"ד, מועד א'
משך הבחינה: 3 שעות, יש לענות על כל השאלות
חומר עזר: מחשבון, 5 דפי נוסחאות, טבלאות סטטיסטיות

פתרון

שאלה 1 (40 נקודות)

לכבוד הקיץ, שוקי הסמנכ"ל החדש של רשת מלונות "אצטרובל" החליט לבחון את הגורמים המשפיעים על מחירי הלינה בשוק תיירות הפנים. לשם כך הוא ערך סקר קצר בו הוא בדק את הקריטריונים הבאים: מה מספר החדרים במלון,האם המלון נמצא בעיר חוף (לדוגמא: תל אביב) ומהי העלות ללילה. להלן הפלטים שהתקבלו:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-28.243	127.564			
	X1	1.587	.558			
	Beach	64.254	143.281			
	X1_Beach	-.987	.578			

a. Dependent Variable: Y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	78082.740	3	26027.580	8.852	.001[b]
	Residual	44103.891	15	2940.259		
	Total	122186.632	18			

a. Dependent Variable: Y

b. Predictors: (Constant), X1, Beach, X1_Beach

א. (10 נק') מתוך הטבלאות, מהו המודל הנאמד? הסביר את המשמעות של כל אחד מהמקדמים שהתקבלו.

כתיבת המודל נכון 4 נק' (אי כתיבת מודל מספרי 2- נק')

$$\hat{y} = -28.243 + 1.587X_1 + 64.254X_2 - 0.987X_1X_2$$

כתיבת משמעות כל מקדם 1.5 נק'

$\hat{\beta}_0$ - החותך במודל עבור בתי מלון בערים ללא חוף

$\hat{\beta}_1$– תוחלת התוספת לעלות ללילה בעבור תוספת חדר יחיד בבתי מלון בערים ללא חוף

$\hat{\beta}_2$– תוחלת התוספת השולית לחותך עבור בתי מלון בערי חוף

$\hat{\beta}_3$– תוחלת התוספת השולית לעלות ללילה בעבור תוספת חדר יחיד בבתי מלון בערי חוף

ב. (10 נק') להלן טבלה של המודלים האפשריים, עבור כל מודל נתונים המדדים R^2_{adj}, SSE ו-AIC השלם את ארבעת הנתונים החסרים בטבלה. הערכים בטבלה

חישוב נכון של כל נתון חסר 2.5 נק' (נימוק ש- R^2_{adj} לא יכול להיות שלילי 2- נק')

Model	R^2_{adj}	SSE	AIC
X1	0.028	112211.003	168.9902
Beach	0.013	113843.774	169.2647247
X1_Beach	-0.059	122172.516	170.6062578
X1, Beach	0.515	52682.399	156.6243566
X1, X1_Beach	0.588	44695.203	153.5004671
Beach, X1_Beach	0.375	67888.557	161.4424921
X1, Beach, X1_Beach	0.567	44103.891	155.2474219

ג. (10 נק') בצע מבחן הבודק האם מיקום המלון בעיר חוף משפיע על עלות הלינה ללילה (בהינתן שמספר החדרים משפיע על עלות הלינה). רשום את ההשערות ובדוק אותן ברמת מובהקות של 5%.

רישום השערות מדויק 2 נק', בחירת F_{cr} נכון 2 נק', חישוב F_{st} 4 נק', מסקנה סופית 2 נק'
ביצוע מבחן t לא התקבל כלל, מבחן F חלקי עבור השערות חלקיות קיבל ניקוד חלקי
יש לשים לב כי משתנה "מספר החדרים" קיים במודל.

$H_0: \beta_2 = \beta_3 = 0$

$H_1: else$

$$F_{st} = \frac{[SSE(RM) - SSE(FM)]/2}{MSE(FM)} = \frac{(112211.003 - 44103.891)/2}{2940.259} = 11.581$$

$$F_{cr} = F_{2,15}^{0.95} = 3.682$$

קיבלנו כי הסטטיסטי גדול מהערך הקריטי ולכן נאמר כי המודל מובהק.

ד. (10 נק') שוקיהציע את המודל הבא:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + \varepsilon_i$$

D_{1i} - משתנה דמי המציין אם המלון בעיר חוף
D_{2i} - משתנה דמי המציין אם המלון בעיר שאינה גובלת בחוף
X_i - מספר החדרים במלון
Y_i - עלות לינה ללילה

הסבר כיצד הוא שונה מהמודל המלא בסעיפים הקודמים. האם על סמך הנתונים הקיימים ניתן לקבוע אם מודל זה יותר טוב או פחות טוב מהמודל המקורי. נמק תשובתך.

רק תשובה מלאה זיכתה במלוא הנקודות (חובה לציין את בעיית המולטי קולינאריות)

המודל שונה בהגדרת משתני הדמה. בנוסף, כעת אין משתנה אינטראקציה, כלומר אין התייחסות לכך שיכולה להיות השפעה שונה של תוספת חדר במלון בעיר חוף למלון בעיר ללא חוף. **הבעיה העיקרית** במודל זה היא העובדה שישנה <u>מולטי קולינאריות</u> עקב הגדרה של 2 משתני דמה. כפי שלמדנו יש להגדיר את מספר משתני הדמה בהתאם למספר הקטגוריות פחות 1, במקרה זה ישנן 2 קטגוריות (עם חוף/בלי חוף) וכל עוד החותך נמצא במודל זה בעייתי.

שאלה 2 (30נקודות)

סוכן בורסה עקב אחרי שער הדולר באגורות (Y) במשך שלושים ימי עסקים (X=1,...,30). הסוכן התלבט האם ניתן להתאים לנתונים שהתקבלו מודל רגרסיה ליניארית. לשם כך הוא חישב מודל רגרסיה בנפרד עבור עשרת הימים הראשונים (X=1,...,10) ועבור עשרת הימים האחרונים (X=21,...,30). התוצאות שהתקבלו בכל אחד משני המודלים מופיעות בטבלה הבאה:

מודל	ימים	חותך	שיפוע	SSE	R^2
A	X=1,...,10	348.2	2.11	118.11	0.76
B	X=21,...,30	346.96	0.61	14.97	0.67

א. (6 נק') האם הקשר בין המשתנה המסביר למשתנה המוסבר מובהק (בר"ב 0.95) בכל אחד משני המודלים הנ"ל (A ו-B)?

טעויות נפוצות:

-הסבר מילולי בלבד, ללא מבחן פורמלי

- שימוש בד"ח לא נכונות או ר"מ לא נכונה

- טעות בשונות של המקדמים או בנוסחאת הסטטיסטי לביצוע המבחן

-טעות בנוסחא לחישוב SXX

- מסקנה סופית לא נכונה- דחיית Ho משמעותה כי המודל מובהק.

נבצע מבחן t עבור כל אחד מהמודלים בנפרד.

$H_0: \beta_1(A) = 0$

$H_1: else$

$$MSE = \frac{118.11}{8} = 14.764$$

$$SXX = \sum_{i=1}^{10} X_i^2 - n\bar{X}^2 = 82.5$$

$$t_{st} = \frac{2.11}{\sqrt{\frac{14.764}{82.5}}} = 4.988$$

$$t_{cr} = t_{8,0.975} = 2.306$$

$H_0: \beta_1(B) = 0$

$H_1: else$

$$MSE = \frac{14.97}{8} = 1.871$$

$$SXX = \sum_{i=21}^{30} X_i^2 - n\bar{X}^2 = 82.5$$

$$t_{st} = \frac{0.61}{\sqrt{\frac{1.871}{82.5}}} = 4.05$$

$$t_{cr} = t_{8,0.975} = 2.306$$

קיבלנו כי הסטטיסטי גדול מהערך הקריטי בשני המודלים, ולכן נאמר כי הקשר מובהק עבור כל אחד מהמודלים.

ב. (6 נק') האם ניתן לבדוק בעזרת הנתונים את הנחת שוויון השונויות? אם כן, מה מסקנת הבדיקה (בר"ב 0.95← *בפועל נאמר להתייחס לר"מ 0.9 על מנת שיוכלו לפתור עם טבלת α=5%F*)? אם לא, מדוע?

טעויות נפוצות:

- טענה כי לא ניתן ניתן לבחון את הטענה.

- הסבר מילולי ללא שימוש במבחן פורמלי.

- ציון המבחן ללא ביצוע המבחן

- ניסוח לא נכון של סטטיסטי המבחן

- שימוש בד"ח לא נכונות
- אי תאימות בין דרגות החופש של הערך הסטטיסטי והערך הקריטי
- אי דחיית H0 למרות שערך הסטטיסטי באיזור הדחייה
- ניסוח לא נכון של הנחות המבחן

נבצע מבחן F ליחס שונויות

$$H_0: \frac{\sigma_A^2}{\sigma_B^2} = 1$$

$$H_1: else$$

$$F_{st} = \frac{14.764}{1.871} = 7.889$$

$$F_{cr(high)} = F_{8,8,0.05} = 3.4381 \; ; \; F_{cr(low)} = F_{8,8,0.95} = \frac{1}{F_{8,8,0.05}} = \frac{1}{3.4381} = 0.291$$

ניתן לראות כי הסטטיסטי נמצא באזור הדחייה, כלומר, הנחת שיוויון השונויות אינה מתקיימת.

ג. (6 נק') האם ניתן לבדוק בעזרת הנתונים את הנחת הנורמאליות? אם כן, מה מסקנת הבדיקה? אם לא, מדוע?

טעויות נפוצות:
- טענה כי ניתן לבדוק הנחה זו (הסקת הנחת הנורמליות על בסיס ערכי X, R^2 או מבחנים פורמליים אחרים ללא ציון מסקנת הבדיקה)
- חוסר ציון הנתונים החסרים במדויק וכי בדיקת הנחת הנורמליות מתבצעת על השאריות.

הנחת הנורמליות של השאריות נבדקת באמצעות השאריות ולכן בהיעדר ערכי התצפיות (או השאריות) לא ניתן לבדוק את הנחת הנורמאליות, שכן יש צורך בערכי התצפיות שהתקבלו בפועל שבעזרתם יהיה ניתן לחשב את השאריות.

ד. (6 נק') לאור ערכי השיפוע המופיעים בטבלה, מה דעתך על הנחות מודל הרגרסיה עבור כל הנתונים (30 ימים)?

טעויות נפוצות:
- הסקת הנחת המודל, מלבד הנחת הליניאריות, על בסיס ערכי X או על בסיס ערכי השיפועים הנתונים
- חוסר התייחסות לקיום הנחת הליניאריות או השלכת קיום הנחה זו על בסיס סעיף א'.
- פירוט חסר

לאור ההבדל המשמעותי בין השיפועים, סביר להניח שהנחת הלינאריות אינה מתקיימת, שכן עבור תחילת החודש השיפוע תלול באופן יחסי לעומת השיפוע המתקבל בשליש האחרון של החודש.

ה. (6 נק') האם ניתן לבסס את התשובה לשאלה ד' באמצעות מבחן סטטיסטי פורמאלי שמסתמך על הנתונים המוצגים בלבד? אם כן, כיצד? אם לא, מדוע?

טעויות נפוצות:
- רישום כי לא ניתן לבצע מבחן פורמלי ללא ציון מבחן מסויים וללא מתן הסבר מדוע לא ניתן לבצע מבחן שכזה.
- הצעת מבחן LOF כמבחן פורמלי אפשרי ללא ביצועו או מתן מענה על כיצד לבצעו

לא ניתן לבסס את התשובה על ידי מבחן פורמאלי מדויק. בסעיף ב' קיבלנו שהנחת שיוויון השונויות של שני המדגמים הללו אינה מתקיימת, ולכן לא נוכל לבצע מבחן אחוד על הפרש

השיפועים. בנוסף, לא נוכל לבדוק האם שיפוע מסוים שווה לשיפוע האחר (כערך קבוע), כיוון שבמבחן כזה אנו "כופים" מהו השיפוע האמיתי, וזה כמובן ערך המתקבל מהמדגם הקיים בלבד.
בנוסף, **לא ניתן לבצע מבחן CHOW שכן מבחן זה אינו תקף כאשר הנחת שוויון שונויות אינה מתקיימת (בהתבסס על התפלגות סטטיסטי המבחן).**

<u>שאלה 3 (30נקודות)</u>

הנח כי עבור 15 תצפיות התבצעו חישובי רגרסיה עם חותך ושלושה משתנים מסבירים X_1, X_2, X_3.

Model		Unstandardized Coefficients B	t
1	(Constant)	-1.0337	-10.124
	X1	19.80696	6.672
	X2	1.078324	2.331
	X3	-0.70597	-4.219

Coefficient[a]

Model		X1	X2	X3
Covariances	X1	8.813	0	0.248
	X2	0	0.214	-0.076
	X3	0.248	-0.076	0.028

a. Dependent Variable: Y

****במהלך המבחן נאמר כי מטריצת *cov(X)* ומטריצת *cov(β)* הינן זהות במקרה הספציפי.**

א. (6 נק') השלם את הערכים החסרים בטבלה (A,B,C).הערכים בטבלה.

A,C על ידי סימטריה במטריצת ה-*cov(β)* שכן *cov(β1,β2)=cov(β2,β1)*וכך הלאה

B חישוב בעזרת טבלת ערכי המקדמים

$$t_{st} = \frac{\hat{\beta}_2}{\sqrt{V(\hat{\beta}_2)}} \rightarrow V(\hat{\beta}_2) = \left(\frac{\hat{\beta}_2}{t_{st}}\right)^2 = \left(\frac{1.078324}{2.331}\right)^2 = 0.214$$

ב. (6 נק') בין שני אלו משתנים מסבירים המתאם הזוגי הוא הגבוה ביותר?

$$\rho = \frac{\mathrm{COV}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

$$\rho_{X1,X2} = \frac{0}{\sqrt{8.813 \cdot 0.214}} = 0$$

$$\rho_{X1,X3} = \frac{0.248}{\sqrt{8.813 \cdot 0.028}} = 0.499$$

$$\rho_{X2,X3} = \frac{-0.076}{\sqrt{0.214 \cdot 0.028}} = -0.982$$

ניתן לראות כי ישנו מתאם זוגי גבוה ביןX_2 ו-X_3

ג. (6 נק') האם מתאם זוגי גבוה בין משתנים מסבירים עלול להעיד על בעיה? איזו?

מתאם זוגי גבוה יכול להעיד על בעיית מולטיקולינאריות, בה יש קשר בין משתנה מסביר אחד למשנהו.

ד. (6 נק') באופן כללי, האם יש גורמים במטריצת Covariances אשר חייבים להיות חיוביים?

הערכים הנמצאים באלכסון חייבים להיות חיוביים, שכן הם מייצגים את האומדים לשונות עבור כל אחד מהמקדמים.

ה. (6 נק') מצא רווח סמך ברמת בטחון של 95% לביטוי $2\beta_1 + 3\beta_2$.

על מנת למצוא רווח סמך (לתוחלת) יש לחשב אומד נקודתי ואת השונות של אומד זה

$$2\hat{\beta}_1 + 3\hat{\beta}_2 = 2(19.80696) + 3(1.078324) = 42.8489$$

$$\hat{V}(2\hat{\beta}_1 + 3\hat{\beta}_2) = 4\hat{V}(\hat{\beta}_1) + 9\hat{V}(\hat{\beta}_2) + 2 \cdot 3 \cdot 2\widehat{COV}(\hat{\beta}_1, \hat{\beta}_2) = 37.178$$

$$t_{15-4,0.975} = 2.201$$

$$42.8489 \pm 2.201 \cdot \sqrt{37.178} = [29.429\ ,56.269]$$

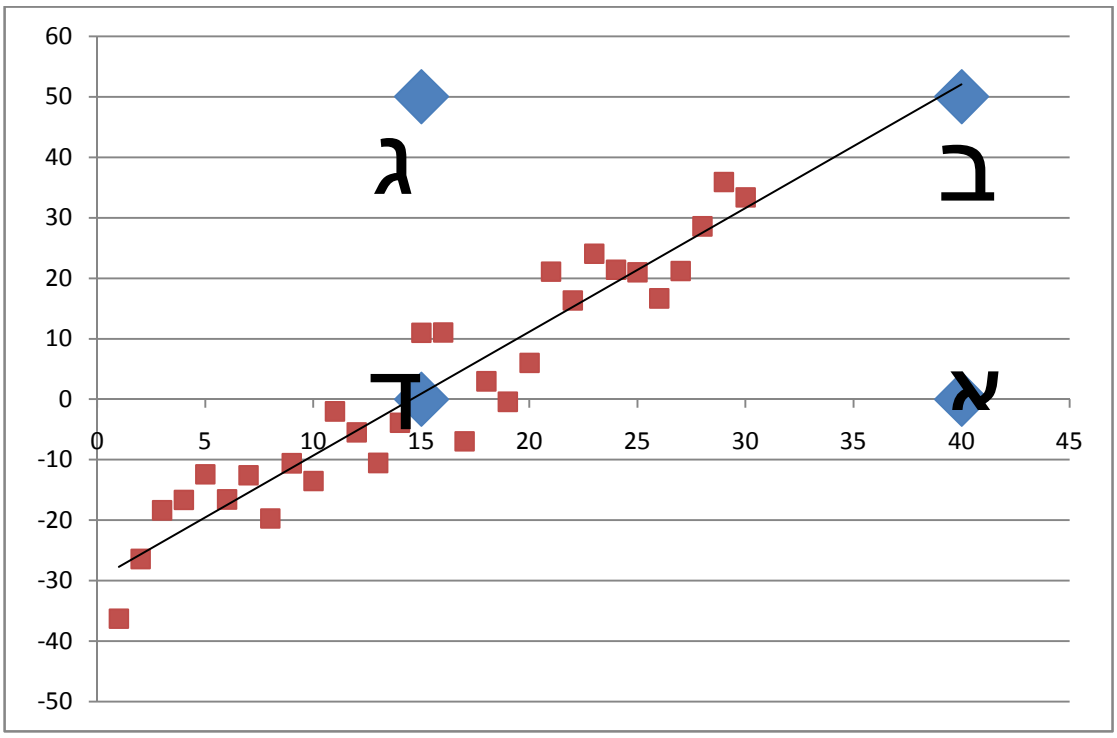
<u>שאלה 4 (בונוס - 20 נקודות)</u>

עבור כל אחד מהצירופים הבאים ציין האם הוא אפשרי או לא. במקרה שהצירוף אינו אפשרי, נמק מדוע. במקרה שהצירוף אפשרי, הצג דוגמא באמצעות תרשים פיזור (אין צורך להציג חישובים נומריים).

א. (5 נק') מדד מינוף (Leverage) גבוה ומדד השפעה (Cook's Distance) גבוה.

ב. (5 נק') מדד מינוף גבוה ומדד השפעה נמוך.

ג. (5 נק') מדד מינוף נמוך ומדד השפעה גבוה.

ד. (5 נק') מדד מינוף נמוך ומדד השפעה נמוך.

פתרון