אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

הנחיות חשובות:
- יש לפתור את כל השאלות
- יש ללוות את דרך הפתרון בהסברים מילוליים קצרים
- ניתן להשתמש בכל הנוסחאות והתוצאות שפותחו או הוצגו בכיתה ואין צורך להוכיחן מחדש אלא אם כן זה נדרש במפורש
-בכל מבחן השערות יש לנסח את ההשערות הנבדקות, לציין מהו סטטיסטי המבחן וכיצד הוא מתפלג, ולכתוב מהו אזור הדחייה (זאת כמובן בנוסף לתוצאת המבחן)
- אם לא צוין אחרת יש להשתמש בר"מ של 5%

תאריך הבחינה: 10.09.2019
שם המרצה: פרופ' הלל בר-גרא ומר איגור רחלין
שם הקורס: מודלים של רגרסיה לינארית
מספר הקורס: 364.1.1091
שנה: 2019 סמסטר: ב' מועד: ג'
משך הבחינה: 3 שעות
חומר עזר: 5 דפי נוסחאות + טבלאות סטטיסטיות + מחשבון

בהצלחה ☺

שאלה מספר 1 (35 נק')

ג'ון בילדינג רוצה לעשות את צעדיו הראשונים בעולם הנדל"ן. הוא אסף נתונים על 506 עיירות שונות בסביבות בוסטון, במטרה להסביר את המחיר החציוני (באלפי דולרים) של בתים באזור המדובר. ג'ון החליט שהמנבא הטוב ביותר למחיר החציוני של הדירה הוא רמת הפשיעה בעיירה (כמות פשעים רשומים לנפש). ג'ון חישב ומצא שהממוצע של רמת הפשיעה הוא 3.613. בנוסף, ג'ון קיבל את הפלט הבא:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-16.957  -5.449  -2.007   2.512  29.800 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.03311    0.40914   58.74   <2e-16 ***
crim        -0.41519    0.04389   -9.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.484 on 504 degrees of freedom
```

א. (7 נק') חשבו והשלימו את טבלת האנובה המתאימה למודל. נסחו את ההשערה המתאימה וענו עליה בר"מ של 5%.

$$SSE = 8.484^2 \cdot (504) = 36277.041$$

$$\hat{V}(\widehat{\beta_1}) = \frac{MSE}{S_{xx}} = 0.04389^2 \rightarrow S_{xx} = 37365.445$$

$$SSR = \widehat{\beta_1}^2 S_{xx} = 6441.157$$

$Source$	SS	$d.f$	MS	F_st
R	6441.157	1	6441.157	89.487
E	36277.041	504	71.978	
T	42718.198	505		

$$\begin{cases} h_0: \beta_1 = 0 \\ h_1: \beta_1 \neq 0 \end{cases} \rightarrow F_{st} = 89.487 \sim F_{1,504} \rightarrow F_{st} > F_{cr}$$

ולכן נדחה את ההשערה, ונאמר שעל סמך הנתונים, ישנו קשר בין רמת הפשיעה למחיר החציוני של הדירה בר"מ של 5%.

ב. (7 נק') חשבו את מדד טיב ההתאמה. האם התוצאה שקיבלתם מתאימה למסקנה שקיבלתם בסעיף א'? הסבירו.

$$R^2 = \frac{SSR}{SST} = 0.15$$

נראה שמדד טיב ההתאמה נמוך למרות מובהקות המשתנה. יכול להיות שישנו קשר אך הוא אינו לינארי, או שהמגמה לינארית אך שונות התצפיות גבוהה, ולכן נותר חלק מהותי שלא ניתן להסביר באמצעות משתנה רמת הפשיעה.

ג. (7 נק') מה הסיכוי שבעיירה מסוימת בעלת רמת פשיעה של 5 פשיעות לנפש, המחיר החציוני לדירה יהיה נמוך מ28 אלף דולר?

$$(\widehat{y_o}|x_0 = 5) = 21.957$$

נשתמש בשגיאת החיזוי כמשתנה הציר

$$P(y_0|x_0 = 5 < 28000) = P\left(e_0 < \frac{28000 - 21957}{\sqrt{\hat{V}(e_0)}}\right)$$

נחשב את האומדן לשונות השגיאה

$$\hat{V}(e_0) = MSE\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right) = 72.124$$

ולכן

$$P\left(e_0 < \frac{28 - 21.957}{\sqrt{\hat{V}(e_0)}}\right) = P(t_{504} < 8.492) = 0.711$$

כלומר על פי הטבלה (לפי ד"ח 120) הסיכוי הוא בין 0.75 ל0.8.

ד. (7 נק') ידוע כי התפלגויות מחירי הדירות בעיירות בסביבות בוסטון הינן בעלת זנב ימני. הסבירו מה יקרה לאומדנים, אם במקום לנבא את המחיר החציוני, ג'ון היה מנבא את המחיר הממוצע.

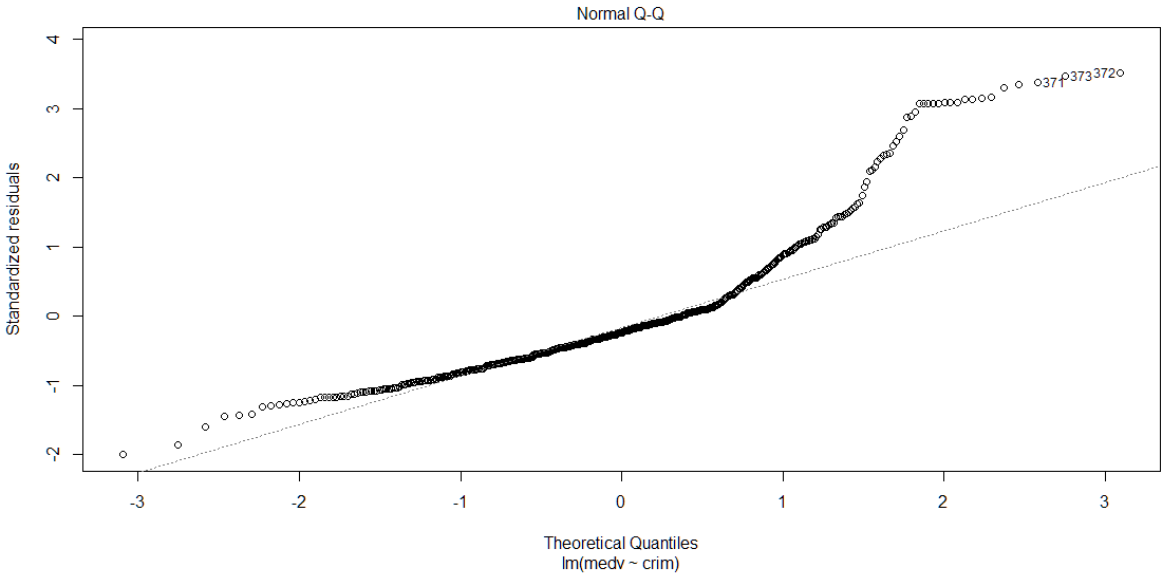
זנב ימני מעיד שהממוצע גדול מהחציון, כלומר כל ערכי המוסבר יגדלו (לאו דווקא בצורה לינארית). היות וערכי המסביר לא ישתנו וערכי המוסבר יגדלו, ערך החותך יגדל. היות ולא

ברור מה יקרה לנקודת הממוצעים (מבחינת ממוצע המוסבר) לא ניתן להגיד בוודאות מה יקרה לשיפוע.

ה. (7 נק') הסבירו את המסקנות מהפלטים הבאים:

```
Shapiro-Wilk normality test

data:  residuals(model1)
W = 0.86696, p-value < 2.2e-16
```



האם המסקנות תואמות? הסבירו.

שני הפלטים באים לבדוק את נורמליות השאריות. לפי מבחן שפירו ווילקס ניתן לראות שדוחים את השערת ה-0, כלומר על סמך הנתונים אינם מתפלגים בצורה נורמלית. הפלט מחזק את מסקנת המבחן, כי ניתן לראות שישנו זנב ימני ארוך מאד להתפלגות.

<u>שאלה מספר 2 (35 נק')</u>

חברו של ג'ון, ג'ימי אפרטמנט, החליט לנבא את המחיר החציוני של הדירה (עבור אותם הנתונים) באמצעות מס' החדרים הממוצע בדירה. ג'ימי קיבל את הפלט הבא:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-23.346  -2.547   0.090   2.986  39.433 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -34.671      2.650  -13.08   <2e-16 ***
rm             9.102      0.419   21.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 6.616 on 504 degrees of freedom
Multiple R-squared:  0.4835,   Adjusted R-squared:  0.4825
F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

א. (7 נק') השוו בין המודלים (של ג'ון ושל ג'ימי) באמצעות שני מדדים שונים.

לפי שונות השגיאות, המודל של ג'ימי (החדש) טוב יותר (כי השונות נמוכה יותר).

היות ושני המודלים בעלי אותה כמות פרמטרים, מדדי הAIC והBIC יניבו מסקנה דומה (כי ההבדל היחיד בין המודלים הוא בפיזור השגיאות).

ביל סקייסקרייפר החליט לשלב את המודלים של ג'ון וג'ימי ולהוסיף משתנה נוסף- אחוז התושבים במעמד הסוציואקונומי הנמוך ביותר. ביל קיבל את הפלט הבא:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.56225    3.16602  -0.809  0.41873
crim        -0.10294    0.03202  -3.215  0.00139 **
rm           5.21695    0.44203  11.802  < 2e-16 ***
lstat       -0.57849    0.04767 -12.135  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.49 on 502 degrees of freedom
Multiple R-squared:  0.6459,   Adjusted R-squared:  0.6437
F-statistic: 305.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

ב. (7 נק') הציעו בעיה אפשרית שהייתה עלולה לצוץ במודל של ביל. הסבירו באמצעות הפלט אם הבעיה באמת קיימת או לא. אם הבעיה קיימת- הציעו דרך לפתרונה.

לפי אופי המשתנים יש חשש למולטיקולינאריות, היות וניתן להניח שישנו קשר בין רמת הפשיעה לבין אחוז התושבים במעמד הסוציואקונומי הנמוך ביותר. אחת מהדרכים להבחין במולטיקולינאריות היא ערכי t נמוכים למרות ערך F גבוה (מה שנובע מהשונות הגבוהה של האומדים). ניתן לראות בנתונים הנ"ל הבעיה אינה קיימת.

ג. (7 נק') בדקו בר"מ של 5% את נחיצות הוספת המשתנים אחוז התושבים במעמד סוציואקונומי הנמוך ביותר ורמת הפשיעה.

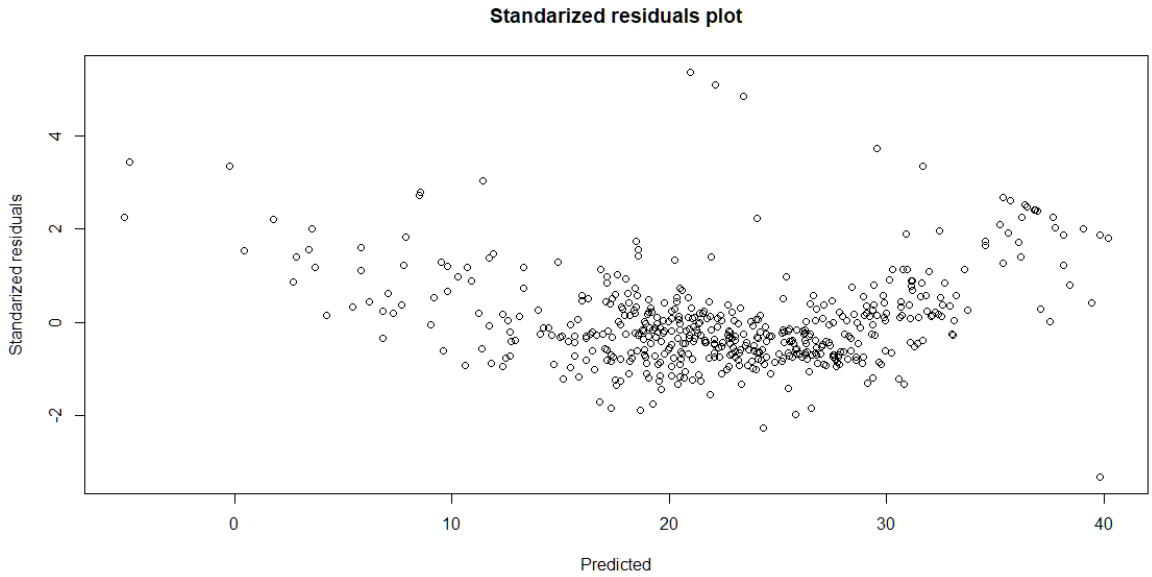
$$\begin{cases} h_0: \beta_2 = \beta_3 = 0 \\ h_1: \qquad else \end{cases}$$

נבצע מבחן F חלקי

$$F_{st} = \frac{(SSE(RM) - SSE(FM))/k}{MSE(FM)} = \frac{\frac{(6.616^2 \cdot 504 - 5.49^2 \cdot 502)}{2}}{5.49^2} = 114.97 \sim F_{2,504}$$

הסטטיסטי גדול מכל קריטי אפשרי ולכן נאמר שעל פי הנתונים, יש צורך בשני המשתנים הנוספים בר"מ של 5%.

ד. (7 נק') מה ניתן ללמוד מהתרשים הבא-



נראה שיש מגמה פרבולית בשאריות, כלומר ישנו קשר ריבועי שלא מודל. כמו כן ניתן לזהות את התצפיות בחלק העליון כחריגות.

ה. (7 נק') איך יראה תרשים הבוקס-קוקס של המודל של ביל? האם יש צורך בטרנספורמציה? נמקו.

לפי הפלט בסעיף הקודם, נראה שיש צורך רק לתקן את בעיית הלינאריות, אותה נתקן באמצעות טרנספורמציה על המשתנים המוסברים. אין צורך בביצוע טרנספורמציה על המשתנה המוסבר, מכיוון שהדבר יכול לגרום לבעיה בשוויון השונויות הקיים.
כל זה מעיד שבתרשים הבוקס קוקס יהיה רווח סמך (כנראה לא גדול) סביב הערך λ=1.

<u>שאלה 3 (30 נק')</u>

נגדיר את וקטור התוחלות $\mu = X\beta$, את התוחלת הממוצעת (השולית) $\bar{\mu}$, ואת סכום ריבועי התוחלות $S\mu\mu = SSP = \sum_{i=1}^{n}(\mu_i - \bar{\mu})^2 = \sum_{i=1}^{n}\mu_i^2 - n \cdot \bar{\mu}^2$. סכום זה מייצג את הפיזור השיטתי במשתנה המוסבר שנובע באופן טהור ישירות מהפיזור במשתנים המסבירים. העזרו בהגדרה של SSP:

א. (21 נק') חשבו את התוחלות של כל אחד ממרכיבי פירוק סכום הריבועים (E(SST), E(SSE), $(E(SSR)$

פיתוח במצגת 3 שקפים 35-39

ב. (9 נק') כתבו שתי מסקנות שנובעות מהתוחלות הנ"ל.

MSE אח"ה ל-σ^2

R^2_{adj} אח"ה לSSP