מעבדה בסטטיסטיקה 52568 התשע״ח

בוחן מסכם – דף הנחיות.

הנחיות: הבוחן כולל 2 שאלות המורכבות ממספר סעיפים. עליכם לענות בטופס הבחינה ובאופן עצמאי על כל השאלות. יש להסביר בקצרה כל תשובה במקום הנתון לה בלבד ובכתב ברור.

חומר עזר: מחשבון.

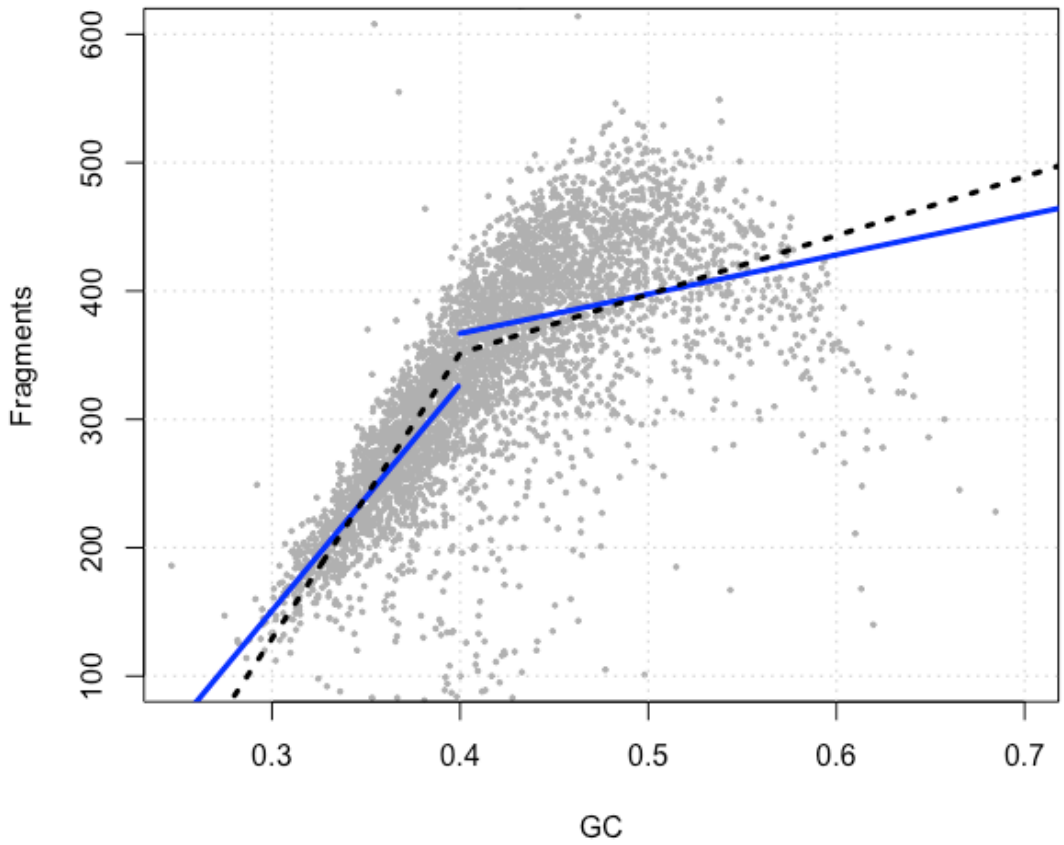
משך הבחינה: שעה

שאלה 1

עדי רוצה לאמוד קו רגרסיה המקשר בין מספר הפרגמנטים בתא וכמות הGC. היא החליטה לאמוד רגרסיה המורכבת ממקטעים ישרים, וכתבה את הנוסחה הבאה:

```
lm(formula = FRAG ~ I(1 * ( GC <0.4)  ) + I(GC * ( GC <0.4) ) +
                    I(1 * ( GC >=0.4) ) + I(GC * ( GC >=0.4)), subset = train_samp)
```

הרגרסיה שאמדה מסומנת בקו מלא בגרף.



א. **[15 נק׳]**

נסמן בX את מטריצת המודל הלינארי (בX 4 עמודות וn שורות). כתבו את השורות של X המתאימות לנקודה עם GC=0.3 ולנקודה עם GC=0.5.

_____

_____

_____

ב. **[20 נק׳]**
ראזי הציע לעשות רגרסיית מקלות שבורים במקום הרגרסיה של עדי. הרגרסיה מתוארת בקו מנוקד. סמנו את התשובה הנכונה והסבירו בקצרה.

ההטיה בריבוע הממוצעת של המודל החדש (ראזי):

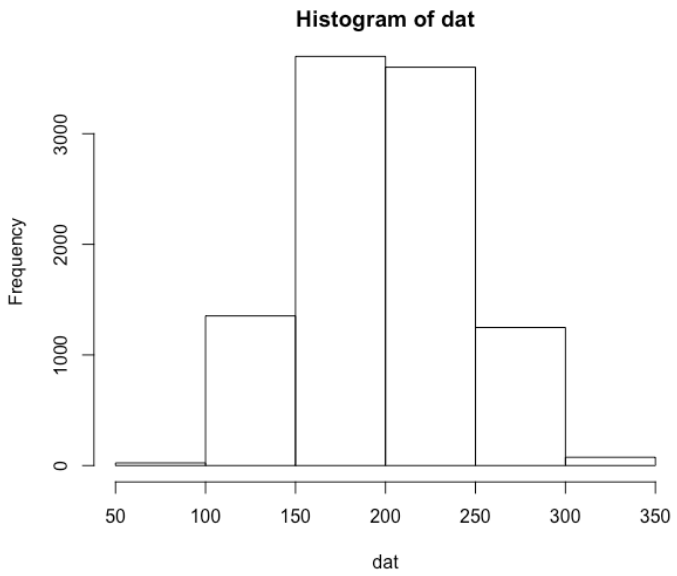
1. גדולה משל המודל של עדי.
2. קטנה מהמודל של עדי.
3. לא ניתן לדעת.

ג. **[10 נק׳]**
הסבירו מדוע נקודת אי רציפות, כמו זו שיש במודל של עדי, היא תכונה לא רצויה עבור פונקציית GC. (רמז – מה יקרה כשננסה לאמוד עותקים?).

שאלה 2

דן רוצה לבדוק את התפלגות הפרגמנטים ב10,000 תאים. הוא משתמש בפונקציית היסטוגרמה ומקבל את הגרף והפלט הבאים:

```
> h
$breaks
[1]  50 100 150 200 250 300 350

$counts
[1]   24 1353 3699 3601 1248   75

$density
[1] 0.000048 0.002706 0.007398 0.007202 0.002496 0.000150

$mids
[1]  75 125 175 225 275 325

$xname
[1] "dat"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```



א. **[15 נקודות]** הציעו שלושה שיפורים שיהפכו את הגרף ליותר אינפרמטיבי.

הצעה 1 _____

_____

הצעה 2 _____

_____

הצעה 3 _____

_____

ב. **[10 נק]** השתמשו בהסטוגרמה כדי לאמוד בקירוב את E[Y] ואת Var[Y].

E[Y] ≈ _____

Var[Y] ≈ _____

ג. **[14 נק]** הניחו שמספר הפרגמנטים מתפלגים פוואסון לא אחיד:

$$\mathrm{Y_k} \sim Poisson(\lambda_k), \quad \lambda_k \sim F,$$

כאשר F התפלגות של התוחלת על פני התאים השונים.

אמדו בקירוב (1) את השונות שאיננה ניתנת להסברה ו(2) את שונות התוחלות [Var[E[Y].

השתמשו באומדנים של התוחלת והשונות שחישבתם בסעיף ב.

ד. **[11 נק]** דן אמד מודל המתאר את השפעת הGC על מספר הפרגמנטים בתא.

כאשר בדק את ממוצע ריבועי השגיאה על <u>נתוני ה train</u> (נתוני האימון):

$$\frac{1}{K_{tr}} \sum_{k \in Train} \left(\mathrm{Y_k} - \hat{f}(GC_k)\right)^2$$

קיבל שגיאה נמוכה בהרבה מהשונות שאיננה ניתנת להסברה (ע״פ מודל פוואסון לא אחיד).

הסבירו בקצרה כיצד זה יתכן, ומה המשמעות של מצב כזה.

טיוטא