שאלה 1

משתני אינטראציה

קובץ נתונים (בדיוני לחלוטין) כולל נתונים על $n = 1{,}000$ ממוצעים עונתיים של קהל במשחקי הבית של קבוצות כדורגל, באלפים (Y). המשתנים המסבירים X_1- תקציב הרכש אותו השקיעה הקבוצה (במילוני יורו), ו-X_2- הליגה בה משחקת הקבוצה (פריימר ליג, לה-ליגה, סרייה א', בונדסליגה).

נתון פלט הרגרסיה של Y על X_1, X_2 כאשר הליגה מקודדת על פי הליגה בה שיחקה הקבוצה:

```
Call:
lm(formula = fans ~ budget * league, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0738  -2.8537   0.0597   2.6556  15.4112 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          42.271730   0.646300  65.406  < 2e-16 ***
budget                0.220877   0.005170  42.721  < 2e-16 ***
leagueLaLiga        -11.959790   0.961079 -12.444  < 2e-16 ***
leaguePremier         8.545799   0.904828   9.445  < 2e-16 ***
leagueSerieA          2.409351   0.952114   2.531 [redacted] *  
budget:leagueLaLiga   0.025686   0.007667   3.350 0.000838 ***
budget:leaguePremier -0.198629   0.007295 -27.229  < 2e-16 ***
budget:leagueSerieA  -0.025224   0.007413  -3.403 0.000694 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.11 on [redacted] degrees of freedom
Multiple R-squared: [redacted],	Adjusted R-squared:  0.8826 
F-statistic:  1074 on 7 and [redacted] DF,  p-value: < 2.2e-16
```

א. כתבו את המטריצה $X \in R^{n\times(p+1)}$ בה השתמשו כדי לאמוד את המודל.

ב. השלימו את הערכים החסרים בפלט (המוסתרים באדום).

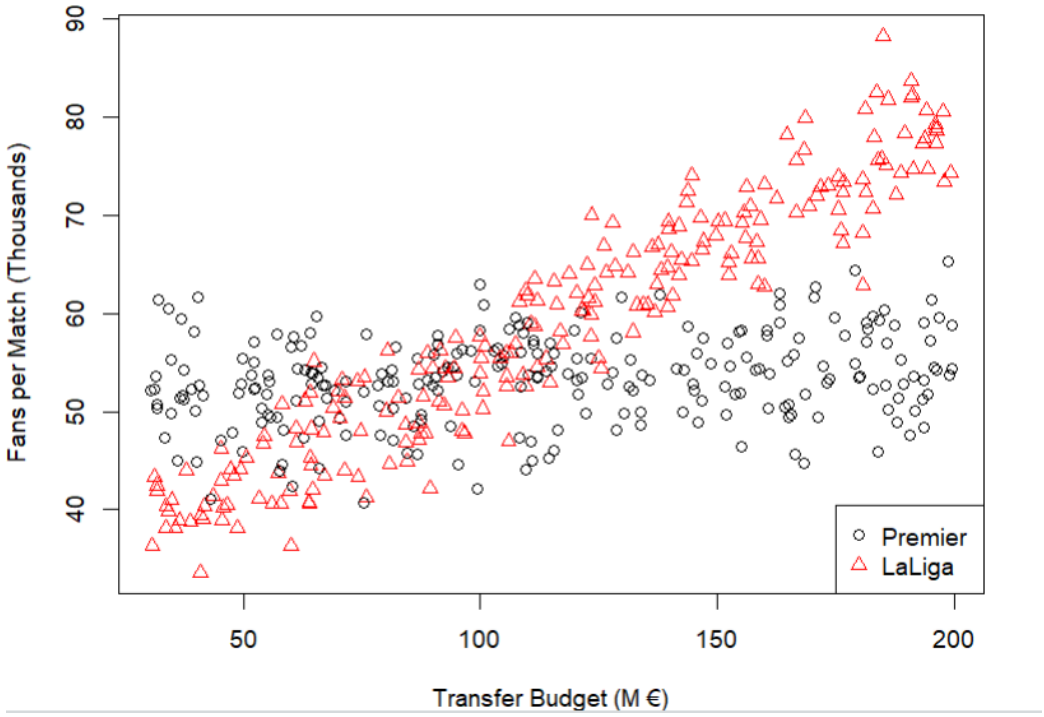
ג. חשבו מרווח חיזוי לממוצע הקהל הביתי בעונה הבאה של קבוצה מהלה-ליגה שתקציב העברות שלה יעמוד על 50 מיליון יורו, ברמת סמך 0.95. השוו את התוצאה למקרה בו הייתם רוצים לחשב רווח סמך לתוחלת ממוצע הקהל הביתי של הקבוצה, והסבירו את המקור להבדל.

ד. מהסתכלות בלתי אמצעית בתרשים הבא בלבד, האם הייתם אומרים כי ישנה עדות לאינטראקציה בין הליגה ממנה הגיעה הקבוצה ובין תקציב הרכש אותו השקיעה? האם ישנה עדות להפרש קבוע בתוחלת כמות האוהדים- לכל רמה של תקציב? נמקו.

ה. נשווה את המודלים עם ובלי אינטראקציה. כלומר:

```
fm1 <- lm(fans ~ budget * league, data = df)
fm2 <- lm(fans ~ budget + league, data = df)
```

האם באופן כללי (ללא קשר לנתונים הספציפיים) ניתן לומר: המקדם של $budget$ יהיה זהה בין שני המודלים. נמקו.



שאלה 1- הנחות המודל והתמודד עם הפרתן

שאלה 1

בשאלה זו ננתח באופן מעשי סטייה מהנחות המודל. במודל מופיע קובץ בשם $ex6_q3_data.csv$. קובץ זה מכיל 3 משתנים מוסברים $y_1,\ y_2,\ y_3$ ומשתנה מסביר x. עבור כל אחד מהמשתנים המוסברים, נרצה לבדוק האם יש סטייה מהנחות המודל הלינארי

$$y_k = \beta_0 + \beta_1 \cdot x + \epsilon, \quad k = 1,2,3 \quad E[\epsilon] = 0$$

לשם כך, עבור כל $y_k,\ k = 1,2,3$ בצעו את השלבים הבאים

א. אמדו מודל לינארי (עבור המשתנים המקוריים).

ב. הציגו היסטוגרמה ו- $qqplot$ (ביחס להתפלגות נורמלית סטנדרטית) של השארית המתוקננת.

ג. הציגו גרף של $(\hat{y}, r)$ ושל (x, r), כלומר גרף של השארית המתוקננת אל מול הערכים החזויים ואל מול המשתנה המסביר.

ד. הסבירו, בהתבססות על התוצאות בסעיפים ב' ו-ג', האם יש אינדיקציה לסטייה מהנחות המודל הלינארי הנורמלי.

ה. במידה והתשובה בסעיף ד' היא חיובית, הציעו טרנספורמציה מתאימה שתתקן את הסטייה מהנחות המודל, חזרו על סעיפים א', ב' ו-ג' עבור המודל המתוקן והראו שהטרנספורמציה אכן סייעה במצב זה.

תזכורת: משתנה מתוקנן: $r_i = Z_{e_i} = (e_i - \bar{e})/\widehat{SD}_e$.