סטטיסטיקה למדעי המחשב – תרגיל בית שבוע 4

שאלה 1 (25 נקודות)

שאלה זאת משתמשת בקובץ הנתונים heights.csv.

א. ציירו תרשים פיזור של גובה (X) ומשקל (Y) מקובץ הנתונים.

ב. הוסיפו את הקו החסין לתרשים בצבע אדום ואת קו הריבועים הפחותים בצבע כחול (מומלץ לממש באופן עצמאי את הקו החסין).

ג. מה ניתן להסיק מהסתכלות על תרשים לגבי הרגישות של כל אחד מהקווים להימצאות תצפיות חריגות?

ד. מצאו את הערכים הבאים: שיפוע קו הריבועים הפחותים, מקדם המתאם בין X ל – Y, אחוז השונות המוסברת. האם לפי נתונים אלו הייתם מסיקים שקיים קשר לינארי בין גובה ומשקל?

ה. חזרו על סעיפים ב' ו-ד' לאחר הסרת תצפית חריגה אחת.

ו. מה ניתן להסיק על ההשפעה של תצפיות חריגות על אחוז השונות המוסברת?

שאלה 2 (25 נקודות)

א. הגרילו באקראי וקטור X באורך 30, מתוך התפלגות נורמאלית עם תוחלת 5 וסטיית תקן 1, בעזרת הפונקציה ()rvs של scipy.stats.norm (הפרמטר loc של ()rvs מגדיר את התוחלת, scale את סטיית התקן ו-size את גודל הדגימה).

ב. צרו את וקטור Y על ידי הכפלת X ב-5 והוספת 2 (Y=5X+2).

ג. מה צפוי להיות המתאם בין X ו-Y , $\hat{r}$? הראו שצדקתם בעזרת חישוב בפייתון.

ד. מה צפוי להיות השיפוע של קו הריבועים הפחותים, $\hat{b}$? הראו שצדקתם.

ה. הוסיפו רעש ל-Y מתוך התפלגות נורמאלית עם ממוצע 0 וסטיית תקן 1 באופן הבא:

```
noise= norm.rvs(loc=0, scale=1, size=30)

Y=Y+noise
```

מהם ערכי $\hat{r}$ ו- $\hat{b}$ כעת?

ו. חשבו את ערכי $\hat{r}$ ו- $\hat{b}$ עבור ערכים שונים של סטיית התקן של הרעש (בין 0.5 ל 10).

ז. צרו גרף פיזור אחד המציג את ערכי $\hat{r}$ כפונקציה של סטיית התקן, וגרף פיזור נוסף המציג את ערכי $\hat{b}$ כפונקציה של סטיית התקן.

ח. מה ניתן להסיק מגרף זה על הקשר שבין הרעש של הנתונים למקדם המתאם, אחוז שונות מוסברת, ולאמינותו של קו הריבועים הפחותים.

שאלה 3 (25 נקודות, 5 נקודות לכל סעיף)

שאלה זאת משתמשת בקובץ הנתונים Age_And_Time.csv בסעיפים א'-ג'. סעיפים ד'-ה' משתמשים בקובץ הנתונים countries.csv.

א. הציגו גרף פיזור לגיל (X) וזמן (Y), מצאו את הקו הלינארי $Y = bX$ לפי שיטת הריבועים הפחותים והציגו אותו בתוך תרשים הפיזור. מהו אחוז השונות המוסברת?

ב. בידקו האם יש צורך בטרנספורמציה והאם יש תצפיות חריגות. הסירו תצפיות חריגות ועשו טרנספורמציה במידת הצורך, נמקו את בחירתכם.

ג. חיזרו על סעיף א' לאחר ביצוע ההתאמות בסעיף ב'. איזה מודל מתאים יותר לנתונים?

ד. עבור נתוני countries.csv, מצאו את קו הריבועים הפחותים המתאים לכל אחד מהמודלים הבאים:

- $Y = \beta_0 + \beta_1 X_1$
- $Y = \beta_0 + \beta_1 X_2$

כאשר $Y = life\_expectancy$, $X_1 = income$, $X_2 = education$. דווחו את השונות המוסברת בכל מודל. מי המודל בעל השונות המוסברת הגדולה ביותר? האם זה מפתיע?

ה. האם דרושה טרנספורמציה של אחד המשתנים? בצעו את הטרנספורמציה המתאימה וחשבו מודלים לינאריים חדשים. האם יש שיפור בשונות המוסברת?

שאלה 4 (25 נקודות, 4 נקודות לכל סעיף)

פונקציית הצפיפות והפה"מ של משתנה מקרי מעריכי מופיעה במצגת של תרגול 4.

א. הוכיחו כי פונקציית הצפיפות של מ"מ מעריכי היא פונקציית צפיפות תקנית לכל פרמטר $\lambda > 0$.

ב. הוכיחו את תכונת חוסר הזיכרון של מ"מ מעריכי. כלומר, הוכיחו כי עבור $X \sim \exp(\lambda)$ מתקיים $P(X > t + s | X > s) = P(X > t)$.

ג. פתחו את הביטויים לשונות והתוחלת של מ"מ מעריכי. כלומר, הראו כי עבור $X \sim \exp(\lambda)$ אכן מתקיים $E[X] = \frac{1}{\lambda}, V(X) = \frac{1}{\lambda^2}$.

ד. הראו כי התפלגות מעריכית סגורה תחת כפל בקבוע חיובי. כלומר, הוכיחו כי עבור $X \sim \exp(\lambda), a > 0$ מתקיים כי aX מתפלג מעריכית. מהו הפרמטר של ההתפלגות? רמז: חשבו את הפה"מ.

ה. הזמן שלוקח להגיע לאוניברסיטה מתפלג מעריכית עם תוחלת של 20 דקות. השיעור מתחיל עוד 15 דקות ויצאנו רק עכשיו, מה הסיכוי שלא נאחר?

בונוס (10 נקודות):

נסתכל על מקדם המתאם המדגמי $\hat{r}$:

$$\hat{r} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}} = \frac{\widehat{Cov}(X,Y)}{\widehat{sd}(X)\widehat{sd}(Y)}$$

מהו שיעור התצפיות שנדרש לשנות, ובאיזה אופן, על מנת ש $\hat{r}$ ישתנה מערך חיובי ממש לערך ששואף ל-0. האם זו תמצית חסינה?