סטטיסטיקה למדעי המחשב (0365-2301) | תרגיל 4

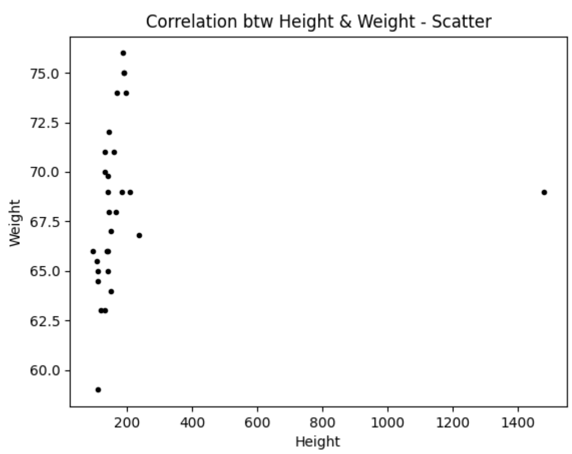
שם: יונתן יהב | שם משתמש: jonathanhuck | ת"ז: 315230417

שם: דור בורשן | שם משתמש: dorbourshan | ת"ז: 315780122

November 21, 2022

שאלה 1

(א)

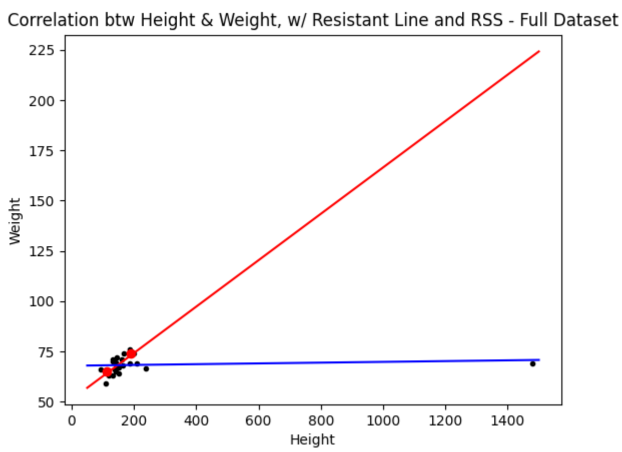


איור 1: תרשים הפיזור

```
# q1
path = "./heights.csv"
df = pd.read_csv(path)
df = df.sort_values("HEIGHT")
heights = df["HEIGHT"].to_numpy()
weights = df["WEIGHT"].to_numpy()
# section  a
pl.scatter(x=heights, y=weights, color="black", marker=".")
```

איור 2: הקוד עבור תרשים הפיזור

(ב)



איור 3: התרשים עם הקו החסין וקו הריבועים הפחותים

```python
# sections b, d, e
MAX_X = 1500 # change to 275 when removing outlier
median = lambda data: data[ceil(data.size / 2) - 1 * (data.size + 1 % 2)]
[lower, mid, upper] = np.array_split(df, 3)
lower_heights = lower["HEIGHT"].to_numpy()
lower_weights = lower["WEIGHT"].to_numpy()
lower_weights.sort()
upper_heights = upper["HEIGHT"].to_numpy()
upper_weights = upper["WEIGHT"].to_numpy()
upper_weights.sort()
x_lower = median(lower_heights)
y_lower = median(lower_weights)
x_upper = median(upper_heights)
y_upper = median(upper_weights)

b_RL = (y_upper - y_lower) / (x_upper - x_lower)
r_is = []
for i in range(heights.size):
    r_i = weights[i] - b_RL * heights[i]
    r_is.append(r_i)
r_is.sort()
a_RL = median(np.array(r_is))
print(a_RL)

x_RL = np.linspace(50, MAX_X, 1000)
y_RL = a_RL + b_RL * x_RL

pl.plot(x_RL, y_RL, "red")
pl.scatter([x_lower, x_upper],  [y_lower, y_upper], color="red")

var_x = np.var(heights)
cov_x_y = np.cov(heights, weights)[0][1]
correlation_x_y = np.corrcoef(heights, weights)[0][1]
print("Correlation coefficient between X and Y:", correlation_x_y)
R_squared = correlation_x_y ** 2 # as we saw in class
print("R^2 = r^2 =", R_squared)
x_avg = heights.sum() / heights.size
y_avg = weights.sum() / weights.size
x_LS = np.linspace(50, MAX_X, 1000)
y_LS = a_LS + b_LS * x_LS
pl.plot(x_LS, y_LS, "blue")

pl.title("Correlation btw Height & Weight, w/ Resistant Line and RSS - Full Dataset")
pl.xlabel("Height")
pl.ylabel("Weight")
pl.show()
```

איור 4: הקוד המשמש אותנו לסעיף (ב) וגם לסעיפים (ד) ו-(ה) שלהלן

(ג)

קל לראות מהתרשים כי קו הריבועים הפחותים מושפע הרבה יותר מתצפיות חריגות מאשר הקו החסין. הדבר הולם את התיאוריה כפי שראינו אותה בכיתה, שכן הראשון נקבע על בסיס שברונים ואילו השני מבוסס מומנטים.

(ד)

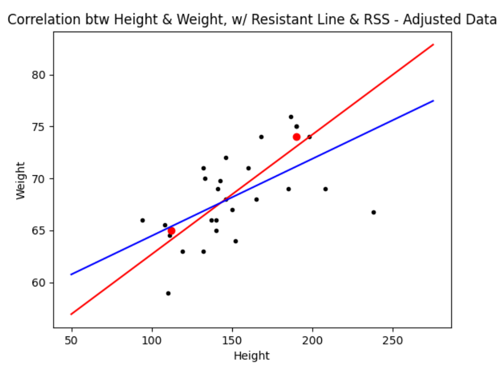
```
Correlation coefficient between X and Y: 0.11163594809565441
R^2 = r^2 = 0.012462584907215647
Sum of squares slope: 0.0019115603603694275
```

איור 5: מקדם המתאם, השונות המוסברת, ושיפוע קו הריבועים הפחותים

לפי נתותנים אלו, היינו מסיקים שלא קיים קשר לינארי בין גובה למשקל. השונות המוסברת קרובה מאוד לאפס, מה שמעיד על מידת התאמה נמוכה של המודל הלינארי.

(ה)

עבור סעיף זה, הסרנו את התצפית (69,1480), שהגובה הרשום בה חריג ביחס ליתר הנתונים. ניתן להעריך שנפלה טעות ברישום התצפית ונוסף אפס מיותר לגובה (זאת אומרת, שהגובה המקורי היה 148 ולא 1480).



איור 6: התרשים עם הקו החסין וקו הריבועים הפחותים כאשר התצפית החריגה מוסרת

```
Correlation coefficient between X and Y: 0.5857442860266674
R^2 = r^2 = 0.34309636861289033
Sum of squares slope: 0.07430831161676091
```

איור 7: מקדם המתאם, השונות המוסברת, ושיפוע קו הריבועים הפחותים לאחר הסרת התצפית החריגה

לפי נתונים אלו, היינו מסיקים שקיים קשר לינארי חלש עד בינוני בין גובה למשקל.

(ו)

מהסעיפים הקודמים, **ניכר שתצפיות חריגות עלולות לפגום במידה רבה באחוז השונות המוסברת של מודל סכום הריבועים הפחותים עבור מדגם.** ראינו לעיל דוגמה מפורשת למקרה בו תצפית חריגה יחידה קירבה אחוז שונות מוסברת לא מבוטל לאפס וכך גרמה לנו להסיק שלא קיים קשר לינארי בעוד שניתן להבחין בקשר לינארי ברור כאשר אותה התצפית מוסרת מהמדגם הנבחן.

שאלה 2

(א)

```python
import scipy.stats as sc
import numpy as np

# q2
# section a
x = sc.norm.rvs(loc=5, scale=1, size=30)
```

איור 8: הקוד להגרלת הוקטור

(ב)

```python
# section b
y = 5*x + 2
```

איור 9: הקוד ליצירת הוקטור Y

(ג)

מכיוון שהוקטור Y התקבל באמצעות הפעלת פונקציה ליניארית על הוקטור X, ישנה תלות מלאה של הראשון בשני, ואפשר להסיק שמקדם המתאם התיאורטי הוא 1.

```python
# section c
corr = np.corrcoef(x, y)[0][1]
print("Empirical correlation coefficient:", corr)
```

איור 10: הקוד לחישוב מקדם המתאם האמפירי

ואכן, בדיקה לפי המדגם שהגרלנו מניבה מקדם מתאם אמפירי קרוב מאוד ל-1:

```
Empirical correlation coefficient: 0.9999999999999998
```

איור 11: מקדם המתאם האמפירי כפי שמתקבל מחישוב בעזרת פייתון

(ד)

קו הריבועים הפחותים אמור להוות מודל לקשר ליניארי ככל שקיים בין המשתנים. מפני שהוקטור Y התקבל באמצעות הפעלת פונקציה ליניארית על הוקטור X, בפרט כזו ששיפועה 5, נצפה שבתור מודל לקשר, קו הריבועים הפחותים יהיה בעל אותו השיפוע.

```
# section d
b_LS = np.cov(x, y)[0][1] / np.var(x)
print("Empirical sum of least squares slope:", b_LS)
```

איור 12: הקוד לחישוב שיפוע קו הריבועים הפחותים

ואכן, בדיקה לפי המדגם שהגרלנו נותנת שיפוע קו ריבועים פחותים אמפירי קרוב מאוד ל-5, כאשר את ההפרש ניתן לייחס לגודל המדגם:

```
Empirical sum of least squares slope: 5.1724137931034475
```

איור 13: שיפוע קו הריבועים הפחותים האמפירי כפי שמתקבל מחישוב בעזרת פייתון

לצורך העניין, בהגדלת המדגם מ-30 ל-10000 נקבל תוצאה אמפירית אף קרובה יותר לתיאורטית שחזינו:

(ה)

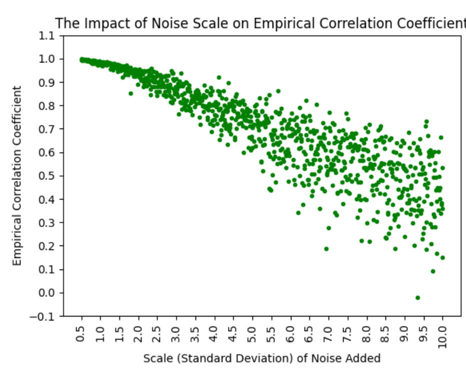
```
Empirical correlation coefficient with noise: 0.987173688482055
Empirical sum of least squares slope with noise: 5.012454724265849
```

איור 14: מקדם המתאם ושיפוע קו הריבועים הפחותים האמפיריים עם הרעש שהוגדר עבור סעיף (ה)

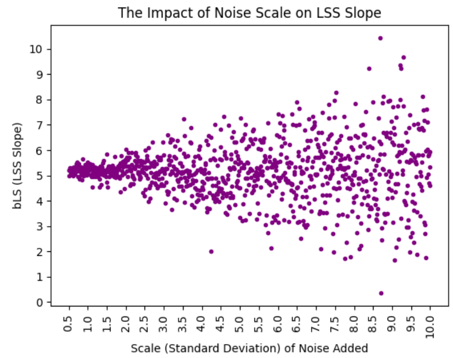
(ו)

ר' קוד מצורף לסעיף (ז) להלן.

(ז)



איור 15:



איור 16:

```
# sections f,g
base_y = 5*x + 2
corrs = []
slopes = []
sds = np.linspace(0.5, 10, 950)
for i in range(950):
    sd = sds[i]
    step_noise = sc.norm.rvs(loc=0, scale=sd, size=30)
    noisy_y = base_y + step_noise
    corrs.append(np.corrcoef(x, noisy_y)[0][1])
    slopes.append(np.cov(x, noisy_y)[0][1] / np.var(x))

pl.scatter(sds, corrs, marker=".", color="green")
pl.title("The Impact of Noise Scale on Empirical Correlation Coefficient")
pl.xticks(np.arange(0.5, 10.5, step=0.5), rotation=90)
pl.xlabel("Scale (Standard Deviation) of Noise Added")
pl.ylabel("Empirical Correlation Coefficient")
pl.yticks(np.arange(-0.1, 1.1, step=0.1))
pl.gcf().subplots_adjust(bottom=0.15)
pl.show()

pl.scatter(sds, slopes, marker=".", color="purple")
pl.title("The Impact of Noise Scale on LSS Slope")
pl.xticks(np.arange(0.5, 10.5, step=0.5), rotation=90)
pl.xlabel("Scale (Standard Deviation) of Noise Added")
pl.ylabel("bLS (LSS Slope)")
pl.yticks(np.arange(0, 11, step=1))
pl.gcf().subplots_adjust(bottom=0.15)
pl.show()
```

איור 17: הקוד לחישוב עבור ערכים שונים של סטיית התקן ויצירת הגרפים שלעיל

(ח)

המסקנה הברורה הינה שככל שסטיית התקן של הרעש שנוסף לתצפיות גדל (כלומר, ככל שמידת הפיזור של ערכי הרעש גדלה), כך נפגעים המדדים להתאמת מודל לינארי למדגם: מקדם המתאם מקבל ערכים רחוקים יותר ויותר מהערך התיאורטי, בהתאם לכך גם אחוז השונות המוסברת המתקבל לפי מקדם המתאם האמפירי, ואמינות קו הריבועים הפחותים יורדת בצורה חדה.

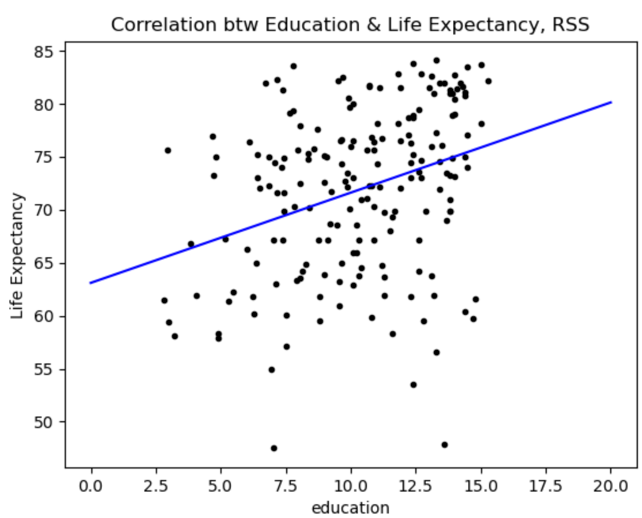
שאלה 3

(ד)

איור 18: $Correlation\ btw\ Income\ and\ Life\ Expectancy$

מתקיים כי

$$Y = \left(2.5 \cdot 10^{-5}\right) X + 67.85$$
$$R^2 = 0.39$$



איור 19: $Correlation\ btw\ Education\ and\ Life\ Expectancy$

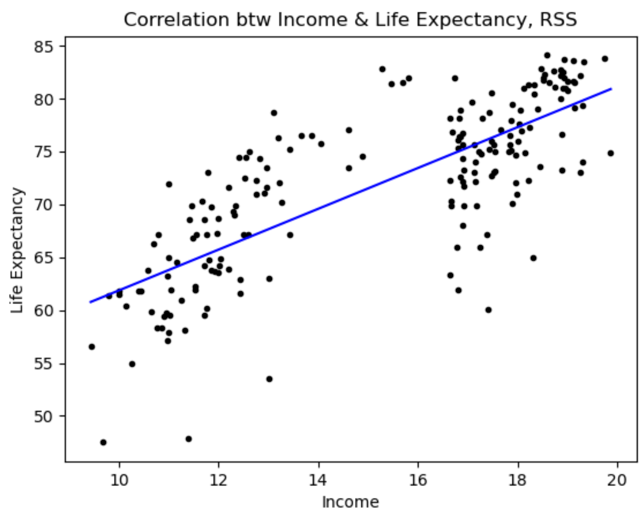
מתקיים כי

$$Y = (0.85)\,X + 63.09$$
$$R^2 = 0.103$$

אנו רואים כי המודל בעל השונות המוסברת הגדולה ביותר הוא מודל ההכנסה.
הדבר אינו מפתיע וזאת משום שאנו משערים כי אנשים בעלי הכנסה גבוהה יותר, יכולים לספק לעצמם רפואה מתקדמת טובה יותר.

(ה)

נבצע טרנספורמציה לוגריתמית ל$Income$



איור 20: $\log_2$ $Correlation\ btw\ Education\ and\ Life\ Expectancy$

מתקיים כי

$$Y = 1.93X + 42.55$$
$$R^2 = 0.588$$

יש שיפור של מעל 40% עבור השונות המוסברת.

שאלה 4

(א)

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

$$f(x) = \lambda e^{-\lambda x}$$

נוכיח כי זוהי פונקצית צפיפות: כלומר אנו נדרשים להוכיח כי

$$\int_{-\infty}^{\infty} \lambda e^{-\lambda x} dx = 1$$

אם כך נחל בחישוב

$$\int_{-\infty}^{\infty} \lambda e^{-\lambda x} dx = \int_{-\infty}^{0} 0 dx + \int_{0}^{\infty} \lambda e^{-\lambda x} dx = \int_{0}^{x} \lambda e^{-\lambda u} du = -\int_{0}^{x} -\lambda e^{-\lambda u} du$$

נסמן $g(x) = e^{-\lambda u}$ אם כך $g'(x) = -\lambda e^{-\lambda u}$
כלומר

$$-\int_{0}^{x} -\lambda e^{-\lambda u} du = -\int_{0}^{x} g'(u)\, du = -g(u)|_{0}^{x} = -e^{-\lambda x} + 1$$
$$\lim_{x\to\infty} -e^{-\lambda x} + 1 = 0 + 1 = 1$$

כנדרש מפונקציית צפיפות.

(ב)

הוכיחו את תכונת חוסר הזיכרון עבור $X \sim \exp(\lambda)$ כלומר הוכח:

$$\mathbb{P}(X > t+s | X > s) = \mathbb{P}(X > t)$$

מההגדרה נחשב את $\mathbb{P}(X > t+s|X > s)$:

$$\mathbb{P}(X > t+s|X > s) = \frac{\mathbb{P}(X > t+s, X > s)}{\mathbb{P}(X > s)} = \frac{\mathbb{P}(X > t+s)}{\mathbb{P}(X > s)} =$$
$$= \frac{\int_{t+s}^{\infty} \lambda e^{-\lambda x} dx}{\int_{s}^{\infty} \lambda e^{-\lambda x} dx} = \frac{\lim_{x\to\infty} -e^{-\lambda x} + e^{-\lambda(t+s)}}{\lim_{x\to\infty} -e^{-\lambda x} + e^{-\lambda(s)}} = e^{-\lambda(t)} = \mathbb{P}(X > t)$$

(ג)

נחשב שונות ותוחלת

$$\mathbb{E}\left[X\right]=\int_{-\infty}^{\infty}xf\left(x\right)dx$$

אם כך נחל בחישוב

$$\int_{-\infty}^{\infty}x\lambda e^{-\lambda x}dx=\int_{-\infty}^{0}0dx+\int_{0}^{\infty}x\cdot\lambda e^{-\lambda x}dx=-\lim_{x\to\infty}\int_{0}^{x}u\cdot-\lambda e^{-\lambda u}du$$

נסמן $g\left(x\right)=e^{-\lambda u}$ אם כך $g'\left(x\right)=-\lambda e^{-\lambda u}$
כלומר

$$-\int_{0}^{x}-\lambda e^{-\lambda u}u\cdot du=-\int_{0}^{x}u\cdot g'\left(u\right)du$$

כעת אינטגרציה בחלקים עבור $\int u\cdot g'\left(u\right)du$

$$x\cdot g\left(x\right)-\int g\left(u\right)du=\int u\cdot g'\left(u\right)du$$

נחשב

$$\int g\left(u\right)du=\int e^{-\lambda u}du=-\frac{1}{\lambda}e^{-\lambda u}+C$$

לכן

$$-u\cdot g\left(u\right)|_{0}^{x}-\frac{1}{\lambda}e^{-\lambda u}|_{0}^{x}=-\int_{0}^{x}u\cdot g'\left(u\right)du$$

$$x\left(e^{-\lambda x}\right)-\frac{1}{\lambda}e^{-\lambda x}+\frac{1}{\lambda}$$

וכעת

$$\mathbb{E}\left[X\right]=-\int_{0}^{\infty}u\cdot g'\left(u\right)du=\lim_{x\to\infty}x\left(e^{-\lambda x}\right)-\frac{1}{\lambda}e^{-\lambda x}+\frac{1}{\lambda}=\frac{1}{\lambda}$$

נחשב כעת

$$Var\left(X\right)=\mathbb{E}\left[X^{2}\right]-\mathbb{E}^{2}\left[X\right]$$

אנו יודעים כי

$$\mathbb{E}^{2}\left[X\right]=\left(\frac{1}{\lambda}\right)^{2}$$

נותר לחשב

$$\mathbb{E}\left[X^{2}\right]=\int_{-\infty}^{\infty}x^{2}\lambda e^{-\lambda x}dx$$

נחשב

$$\int_{-\infty}^{\infty}-\lambda e^{-\lambda x}\cdot x^{2}dx$$

נסמן $g\left(x\right)=e^{-\lambda u}$ אם כך $g'\left(x\right)=-\lambda e^{-\lambda u}$ ונסמן $f\left(x\right)=x^{2}$ וכן $f'\left(x\right)=2x$ אזי מאינטגרציה בחלקים

$$x^{2}\cdot g\left(x\right)-\int2u\cdot g\left(u\right)du=\int u^{2}\cdot g'\left(u\right)du\qquad(\#)$$

חישבנו ממקודם את $\mathbb{E}\left[X\right]=\int_{-\infty}^{\infty}\lambda u\cdot g\left(u\right)du$ ולכן $\int_{-\infty}^{\infty}2u\cdot g\left(u\right)du=\frac{2}{\lambda^{2}}$ ומתקיים כי

$$\lim_{x\to\infty}x^{2}e^{-\lambda x}=0$$

אזי נותר להכפיל במינוס את (#) ונקבל

$$\mathbb{E}\left[X^{2}\right]=-u^{2}\cdot g\left(u\right)|_{0}^{\infty}+\frac{2}{\lambda^{2}}=\frac{2}{\lambda^{2}}$$

ולכן

$$Var\left(X\right)=\mathbb{E}\left[X^{2}\right]-\mathbb{E}^{2}\left[X\right]=\frac{2}{\lambda^{2}}-\left(\frac{1}{\lambda}\right)^{2}=\left(\frac{1}{\lambda}\right)^{2}$$

(ד)

נראה שאם $X\sim\exp\left(\lambda\right)$ אז aX גם מתפלג מעריכית.

נחשב תחילה את פה"מ

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{\infty}^{x} \lambda e^{-\lambda u} du = \lambda \int_{\infty}^{x} e^{-\lambda u} du = -e^{-\lambda u}|_0^x = 1 - e^{-\lambda x}$$

יהיי $a > 0$ נחשב

$$\mathbb{P}(aX \leq x) = \mathbb{P}\left(X \leq \frac{x}{a}\right) = 1 - e^{-\frac{\lambda}{a}x}$$

נסמן $\mu = \frac{\lambda}{a}$ אם כך

$$\mathbb{P}(aX \leq x) = \mathbb{P}\left(X \leq \frac{x}{a}\right) = 1 - e^{-\mu x}$$

מתקיים כי $aX \sim \exp(\mu)$ כיוון שפה"מ קובעת התפלגות.

(ד)

הזמן שלוקח להגיע לאוניברסיטה מתפלג מעריכית עם תוחלת של 20 דקות. השיעור מתחיל עוד 15 דק ויצאנו עכשיו, מה הסיכוי שלא נאחר.
נסמן X זמן שיקח לנו להגיע לאוניברסיטה. מתקיים כי $X \sim \exp(\lambda)$. אנו יודעים כי

$$\mathbb{E}[X] = \frac{1}{\lambda} = \frac{1}{3}$$

אם כך מתקיים כי $\lambda = 3$
אנו רוצים לחשב את ההסתברות הבאה:

$$\mathbb{P}(X \leq 15) = 1 - e^{-3 \cdot \frac{1}{4}} = 0.527$$