**תזכורת** - התאמה עד כה

הקדמה בסטטיסטיקה
הצגה על הסתברות בדידה
הגדרת התפלגות אמפירית
הצגה על התפלגות מותנית, משותפת, גזרתית, חוק ביס ונוסחאות ההתפלגות השולית.
שינוי נתונים: קטגוריאלי, מספרי בדיד

נתונים קטגוריאליים

הדגמה גרפית והסקה מהן
הפרדיקס של סימפסון

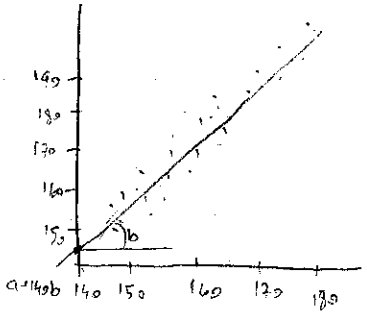
נתונים מספריים חד מימדיים (משתנה אחד).

הדגמות גרפיות: היסטוגרמות, התפלגות גרפן
טרנספורמציות.
סיכומים מספריים - תיאוריות.
הגדרתן: סדרנים (סטטיסטי הסדר), שכיחים ושכיחויות.
תמציות למיקום, לפיזור וצורה.
תכונות של תמציות ובחירה בינהן: חסינות, אינוריאנטיות.

נתונים מספריים רב מימדיים (כמה משתנים יחד).

הדגמות גרפיות ומטרתן (Gapminder)
חשיבות משתנה תלוי מיסבר ומשתנים מסבירים
שיר מתמטי של הקשר לינארי בין משתנה תלוי ומשתנה מסביר יחיד.
קו ישר.
קו רגרסיה בחיתוך.

**הגדרה** - תזכורת על חוקים
- תמציות גירות הקו והחיתוך שלו.
- הרחבתו טרנספורמציות לינאריות.
- רגרסיה בחיתוך עם הרבה משתנים מסבירים.




יש לנו מדגם של צמדות גובה ובנים
$(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)$

$x_i$ - גובה האב ה-i
$y_i$ - גובה הבן ה-i

(*) המטרה שלנו הוא למצוא קו ישר שיתאר נאמנה את הקשר בין המשתנים שלנו.

**קו הרגרסיה הבחיתוך**

$$RSS(a,b) = \sum_{i=1}^{n} (y_i - a - b x_i)^2$$ ← Residual Sum of Squares

קו הרגרסיה הבחיתוך מוגדר ע"י $a_{LS}$, $b_{LS}$ שמביאים למינימום את $RSS(a,b)$.

הרגרסיה בשיטת שפורר:

$$a_{LS} = \bar{y} - b_{LS}\bar{x} \qquad b_{LS} = \frac{Cov_{emp}(x,y)}{Var_{emp}(x)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

מסקנה מעניינת: הנקודה $(\bar{x},\bar{y})$ תמיד נמצאת על קו הרגרסיה האמיתי.

נסתכל על הנתונים הסקטיביים (איך נראים) הבאים:

$\bar{x} = 170$ $\qquad a_{LS} = 56\,cm$

$\bar{y} = 175$ $\qquad b_{LS} = 0.7$ ← (מראים לגבה מנין שהיה קשר בין גובה השם לגובה האב).

(*) נזכר בבעיה שהיינו בשיטת הקודם. אנו רואים ששיטת הנתונים הקיבוציים בעצם משבים מאיד על קו הרגרסיה האמיתי איך נשפיע נגע על הקו האמיתי.

<u>מדדי איכות ההתאמה</u>

הסדרה המוגבקשת: $RSS_{LS} = \sum_{i=1}^{n}(y_i - a_{LS} - b_{LS}x_i)^2$ (*) הטעות הבסדרה לא היא הגלגל שלה הסקטיבה.

אפשרות אחרת: ניקח את $b_{LS}$ במדד לאיכות התאמה. (*) גם כאן המדד מושפע מאיות מהסקטיבה.
"הבעיה": $|b_{LS}|$ גדול מכאן השפעה חלקה של $x$ על $y$.

אפשר להתייחס לבעיה בצורה שונה:
"מהן תכונות מדדית טובה לסימן הקשר בין $x$ ל-$y$?"

<u>מדד שנדון:</u> מקדם המתאם של פירסון (Pearson)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

אמתנו $\qquad b_{LS} = \frac{Cov(x,y)}{Var(x)}$

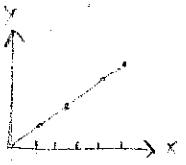
$$r(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)\cdot Var(y)}}$$

$$b_{LS} = r_{xy}\cdot\frac{sd(y)}{sd(x)}$$

תכונות $r$:

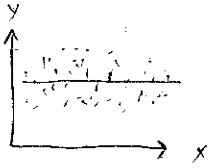
1. $r(x,y) = r(y,x)$
2. אם $b>0$ אז $r(x,a+bx) = r(x,y)$ , $r(a+bx,y) = r(x,y)$
   אם $b<0$ אז $r(a+bx,y) = -r(x,y)$
3. $r(a+bx,y) = \begin{cases} 1 & b>0 \\ 0 & b=0 \\ -1 & b<0 \end{cases}$
4. $-1 \le r(x,y) \le 1$

# סטטיסטיקה למבחן המתאם - שיעור 4



ניקח מדגם $X_1, \dots, X_n$ , $Y_i = a + bX_i$

(*) כלומר $r=1$ אם יש התאמה משלימה בין הנקודות לקו.



אם $r=0$ , גם $b_{LS}=0$ . הסבר הגיוני הוא כיון שהקורלציה האמפירית הוא מדד לקשר לינארי בין $x$ ו- $y$ . במקרה כזה נתן את הבלם שמשמאל (לדוגמה)

יכול להיות מצב שבו יש קו שמתאר באופן מושלם את הנתונים גם כאשר שיפוע הקו הוא 0, למרות המקרה כאשר $y=\text{const}$ .

## $r^2$ ומשמעותו

נזכיר: $\hat{y}_i = a_{LS} + b_{LS}x_i$

אמר:

$$r^2(x,y) = 1 - \frac{\overbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}^{RSS(a_{LS},b_{LS})}}{\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{Var_{am}(y)}} = \frac{Var(y) - RSS(a_{LS},b_{LS})}{Var(y)}$$

**אינטרפרטציה המיסבר ת: $r^2$**

אם $b_{LS}=0$ אזי (מתמטית): $r = r^2 = 0$ ;
סטטיסטית קו רגרסיה באמצעות אינפורמציה לא מסביר כלום על השונות של $y$ .

$$r^2 = \frac{\left(\sum(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2} = \frac{\left(\sum x_iy_i - n\bar{x}\bar{y}\right)^2}{\left(\sum x_i^2 - n\bar{x}\right)\left(\sum y_i^2 - n\bar{y}\right)} \rightarrow$$

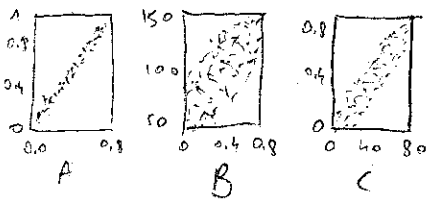
נוסחה לא נדרשת

**בסיכום לגבי שאלה מהמבחן:** בשאלה הקודמת 3 תרשימים (למטה) יש לדרג את ערכי $b_{LS}$, $r$, $r^2$.

הדירוג של $r$ יהיה $A, C, B$ (לפי הפיזור של הנקודות מהקו).

הדירוג של $b_{LS}$ יהיה $C, A, B$ (נזכר כיון שבי $B$ הסקאלה של $y$ גדולה בהרבה מהסקאלה של $x$, לעומת $A$ עם סקאלות דומות ולבסוף $C$ עם סקאלה של $x$ הגדולה מ- $y$).

הסדר של $r^2$ יהיה זהה לשל $r$ כי ערכי $r$ חיוביים. בעיקרון "ליגו" (כלומר לתחום ל- $r$), נתפוס את רמת הפיזור הגדול.




הגרפים הנתונים בשאלה:

**תכונות ותושלת $r^2$**

$r^2$ הוא מדד לגודל המודל.

שבת רגוי להשתמש בו להשוואה רק בין מודלים שהתאמו לנ' נתונים. ← תשלום בפתיחה

מסבר לבבר, $r^2$ רגיש לרמת הנתונים ולטווח הדגם (למשל, עם 2 נקודות, $r^2=1$ תמיד).

# סטטיסטיקה למשי המחשב - שיעור 4

מסתכלים על הקו $\hat{y}$: "קירוב השגיאות מאמצע שיתנו על טרנספורמציות"

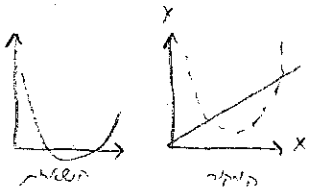
## נניח שגיאות וטרנספורמציה לינארית

$v_i = y_i - \hat{y}_i$ (נמצאר $\hat{y}_i = a_{LS} + b_{LS} x_i$).

אם השגיאות הוא "סדר" או "מבנה" אז הקשר בין x ל-y בכלל אינו לינארי.

דוגמה נניח $y_i = x_i^2$

אם ננסה למצוא $\hat{y}_i = a_{LS} + b_{LS} x_i$ אז יקבל קו שלא מתאר את המידע היטב.




לעומת זאת, אם נבדוק $z_i = x_i^2$ ונתאים קו ישר $\hat{y}_i = a_{LS} + b_{LS} z_i$, אז יקבל התאמה מושלמת כאשר $a_{LS} = 0$ וכן $b_{LS} = 1$.

(*) הרי שהשגיאה ואחר דומה לעצמו, כך הקו מתאר טוב ומתאר את הנתונים.

## מה רואים ב Gapminder?


## בלוגו אדפטיבי הסתכלות על שגיאות

א. אם לשגיאות קורה קטנה (בטווח כלפי מטה) אז יש לנסות טרנספורמציה $x^p$ עבור $p<0$ או $y^q$ עבור $q<1$ אם $b_{LS}<0$.

לעומת זאת, אם $b_{LS}>0$ אז נשתמש $x^p$ עבור $p<-1$ (*) או $y^q$ עבור $q<0$.

ב. אם לשגיאות קורה הטווח (בטווח כלפי מעלה) אז נסה $x^p$ עם $|p|<1$ או $y^q$ עם $q>1$.

## רגרסיה רב משתנית

נסתכל על הנתונים $(x_1,y_1),\dots,(x_n,y_n)$ אל כאשר $x_i \in \mathbb{R}^p$ $p \geq 1$ מספר משתנים מסבירים.

נכתוב: $x_i = (x_{i1},\dots,x_{ip})$ (וקטור שורה).

המודל שנרצה להתאים $\hat{y}_i = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ (כאשר $\beta_0,\dots,\beta_p$ הם פרמטרים שנרצה להתאים).

הקריטון RSS: $RSS(\underline{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \dots - \beta_p x_p)^2$ $\quad \underline{\beta} = (\beta_0,\beta_1,\dots,\beta_p)^T$

נסמן לכתוב מטריצה - וקטורי:

$$(*) \quad RSS(\underline{\beta}) = \|\underbrace{y - X\underline{\beta}}_{\text{וקטור מטריצה}}\|_2^2$$

נורמה רבועית אוקלידית של וקטור

$$y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$$

$$X_{n\times(p+1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{p1} & x_{p2} & \dots & x_{pp} \end{pmatrix}$$

<u>הבעיה שלנו רוצים לפתור:</u> למצוא $\hat{\beta}$ שמביא למינימום את $RSS(\beta)$.

<u>פתרון:</u> נגזור את $(*)$ מטריציונית ונשווה ל-0.

$$\frac{\partial}{\partial\beta}(*) = 2(X^Ty - X^TX\hat{\beta}) = 0$$

$$\Rightarrow X^TX\hat{\beta} = X^Ty \Rightarrow \boxed{\hat{\beta} = (X^TX)^{-1}X^Ty}$$

פתרון המשוואה הנורמלית ש"ל הם מספרי מישתנים

נבדוק עבור $p=1$ קיבלנו קודם: $a_{LS} = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ $\quad b_{LS} = \hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$

נבדוק את התוצאה המטריציונית עבור $p=1$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad X^TX = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

$$(X^TX)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{pmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \quad X^Ty = \begin{pmatrix} n\bar{y} \\ \sum x_iy_i \end{pmatrix}$$

$$(X^TX)^{-1}X^Ty = \frac{1}{n\sum(x_i-\bar{x})^2}\begin{pmatrix} n\bar{y}\sum x_i^2 - n\bar{x}\sum x_iy_i \\ -n^2\bar{y}\bar{x} + n\sum x_iy_i \end{pmatrix}$$

$$\hat{\beta}_0 = ?$$

$$\hat{\beta}_1 = \frac{n\sum x_iy_i - n^2\bar{y}\bar{x}}{n\sum(x_i-\bar{x})^2} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \quad \checkmark$$

<u>נושאי הרצאה ברגרסיה</u>

נתונה רגרסיה עם נשתנים קטגוריאלים
X הקטגוריאלי - הדגלים מספרים,
Y קטגוריאלי - תחום שונה הלוגיסטי, בבעיות רבות
הוא יתנהג יותר ברגרסיה בינארית.

מעבר לרגרסיה לינארית:
רגרסיה לא לינארית
שינוי בקריטריון שחרים מ- RSS.
תיאור קשרים מורכבים בעזרת מתקלות

הסקה סטטיסטית על תקינות הרגרסיה.

## משתנים מקריים רציפים

נזכר בהגדרת התפלגות ומ"מ:

מ"מ $\Omega \subset \mathbb{R}$. במקרה הבדידי הגדרנו $F$ פונ' המתברר: $\sum_{\omega\in\Omega} F(\omega)=1$ $F(\omega)\geq 0$ $\forall \omega\in\Omega$

במקרה הרציף, נקח את $\Omega\subset\mathbb{R}$ למשל לא את כולו. המקרים הרלוונטיים לנו:

א. $\Omega=\mathbb{R}$

ב. $\Omega=[0,\infty)$ או $\Omega=(-\infty,0]$

ג. $\Omega=[a,b]$ $\mathbb{R}\ni a<b\in\mathbb{R}$.

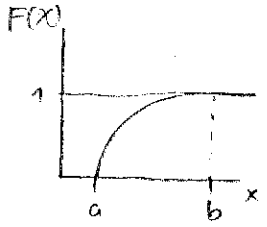
$\Omega$ הוא קטע סגור או פתוח בין $a$ ל- $b$

נגדיר פונקציית התפלגות מצטברת $F$ עם התכונות:

א. $F(a)=0$

ב. $F(b)=1$

ג. $F$ פונקציה לא יורדת.




נאמר שמ"מ $X$ שאינו מקרי מ $\Omega$ בעבר מתפלג לפי $F$ (ונסמן $X\sim F$) אם $\forall x\in\Omega$ מתקיים $Pr(X\leq x)=F(x)$.

$\Leftarrow$ עבור $a,b\in\Omega$ מתקיים $Pr(a\leq X\leq b)=F(b)-F(a)$.

נאמר ש- $F$ היא פונקציית ההס' המצטברת של $X$ ונסמן לפעמים $F_X$.

נגדיר פונקציית צפיפות $f_X(x)=\frac{dF_X(x)}{dx}$

תכונות צפיפות

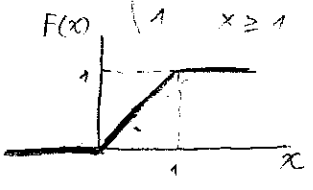
א. $f_X(x)\geq 0$ $\forall x\in\Omega$

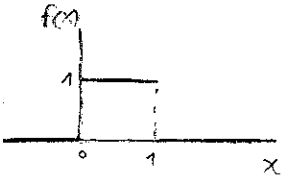
ב. $\int_a^b f_X(x)=1$ הוכחה: $\int_a^b f(x)dx=F(x)\Big|_a^b=1$

דוגמאות:

1. משתנה מקרי רציף שאינו סטנדרטי $X\sim U(0,1)$,

$\Omega=[0,1]$ $f_X(x)=\begin{cases}1 & 0\leq x\leq 1\\ 0 & \text{אחרת}\end{cases}$ $F_X(x)=\begin{cases}0 & x\leq 0\\ x & 0\leq x\leq 1\\ 1 & x\geq 1\end{cases}$




2. מ"מ שאינו באינו $[a,b]$ בשלב.

$f_X(x)=\begin{cases}0 & x<a\\ \frac{1}{b-a} & a\leq x\leq b\\ 0 & x>b\end{cases}$ $F_X(x)=\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a\leq x\leq b\\ 1 & x>b\end{cases}$