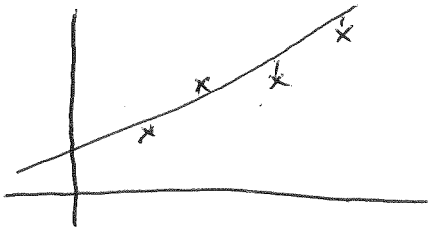
שעור 2 מבוא לסטטיסטיקה והסתברות / ד"ר דוד סרף

למציאות קשר מתימטי בין המשתנה התלוי לבלתי תלויים מבצעים פעולה הנקראת רגרסיה. אם ההתאמה היא לקו ישר הרגרסיה היא ליניארית. הרגרסיה נותנת את מקדמי הפונקציה הטובים ביותר ואת מידת ההתאמה (מקדם המתאם), המראה באיזו מידה הנקודות מטבלת המדידה קרובים לקו.

הרגרסיה המקובלת נעשית בשיטת "הריבועים הפחותים" LSQ (Least Squares). מתפשים את מקדמי הקו אשר ריבועי הסטיות ממנו קטנים ביותר. מקובל למצוא את המרחקים האנכיים



למערכת הצירים, ולא למרחקים הבאומטריים מן הקו עצמו, בגלל שישאר אותנו ה-x גם עבור הנקודה הנסיונית וגם עבור הקו.

א. קו ישר מסוג $y=\beta_1 x$, (העובר דרך הראשית).

מתפשים ערך מינימלי לריבועי הסטיות של הנקודות y_i מן הקו. כדי למצוא את המקדם β_1 הטוב ביותר יש לגזור את הביטוי לפי β_1 ולהשוותו לאפס.

$$\min \sum(\beta_1 x_i - y_i)^2$$

$$2\sum(\beta_i x_i - y_i)\cdot x_i = 0$$

(יש לשים לב x_i, y_i הם ערכים נתונים ולא משתנים)

$$\sum \beta_1 x_i^2 = \sum x_i y_i$$

$$\beta_1 \sum x_i^2 = \sum x_i y_i \quad \Rightarrow \quad \beta_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

לדוגמא: נתונות הנקודות למצוא תרגיל מס 1.

x	1	2	3	4
y	3	5	6	8

את משוואת הקו העובר דרך הראשית

$$\beta_1 = \frac{1\cdot 3 + 2\cdot 5 + 3\cdot 6 + 4\cdot 8}{1^2+2^2+3^2+4^2} = \frac{63}{30} = 2.1$$

לפיכך משוואת הקו תהיה $y = 2.1x$

ב. קו ישר מסוג $\underline{y = \beta_0 + \beta_1 x}$. בדרך כלל אין יודעים את הקו הנייב לעבור דרך הראשית ולכן לוקחים קו כללי שבו β_0, β_1 יש לחפש β_0, β_1 שרבועי הסטיות מן הקו יהיו מינימליים.

מוצרים לפי המשתנים β_0 ו β_1 ומקבלים 2 משוואות ב 2 נעלמים ומהם מחלצים את β_0 ו β_1 נתן כאן את תוצאות הגזירה ~~המשוואות~~.

מינימום $\sum(\beta_0+\beta_1 x_i - y_i)^2$

$$\text{I.}\quad n\beta_0 + \beta_1 \sum x_i = \sum y_i \Rightarrow \beta_0 + \beta_1 \frac{\sum x_i}{n} = \frac{\sum y_i}{n}$$

$$\text{II}\quad \beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \Rightarrow \beta_0 \frac{\sum x_i}{n} + \beta_1 \frac{\sum x_i^2}{n} = \frac{\sum x_i y_i}{n}$$

המשוואות הקנוניות של הרגרסיה

המשוואות הקנוניות של הרגרסיה מספיקות למטלות המקדמים, אבל מצריכות יצירה של כל הנקודות הנסיוניות. ניתן להפוך את המשוואות לסטטיסטיות באמצעות ממוצעים ושונויות. בעבור קבלנו בטויים לשונות, נגדיר ערכי שונות משותפת לב המשתנים הנקראת covariance

$$Cov(x,y) = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y} = \sum \frac{(x_i-\bar{x})(y_i-\bar{y})}{n}$$

טרנספורמציה של הבטוי לשונות המשותפת ע"י טפול אלגברי בבטוי הימני

הגדרת שונות משותפת

תרגיל: הוכח את הטרנספורמציה הנ"ל לפרק בה ראינו כי

$$VAR(x) = \sum (x_i - \bar{x})^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$$

טפולס אולגברי במשוואות הקנונית של הרגרסיה, תוך שימוש במושגי השונות של x $VAR(x)$, השונות המשותפת $Cov(x,y)$ והממוצעים יגיעו 2 משוואות סטטיסטיות, המקבילות למשוואות הקנוניות

I. $\bar{y} = \beta_0 + \beta_1 \bar{x}$ נקודת הממוצעים $(\bar{x},\bar{y})$ יושבת על הקו

II. $\beta_1 = \frac{Cov(x,y)}{VAR\,x}$

x	1	2	3	4
y	3	5	6	8

לדוגמא: נתונה טבלת הנתונים תרגיל מס 2

למצוא את משוואת הקו $y = \beta_0 + \beta_1 x$

נחשב $\bar{x} = 2.5$, $\bar{y} = 5.5$,

$$VAR(x) = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{30}{4} - 2.5^2 = 1.25$$

$$Cov(x,y) = \frac{\sum x_i \cdot y_i}{n} - \bar{x}\bar{y} = \frac{63}{4} - 2.5 * 5.5 = 2$$

לכן $\beta_1 = \frac{Cov}{VAR\,x} = \frac{2}{1.25} = 1.6$

$\beta_0 = 1.5 \Leftarrow 5.5 = \beta_0 + 1.6 * 2.5$

ולכן משוואת הקו $y = 1.5 + 1.6x$

<u>בטוי עבור מקדם המתאם R:</u>

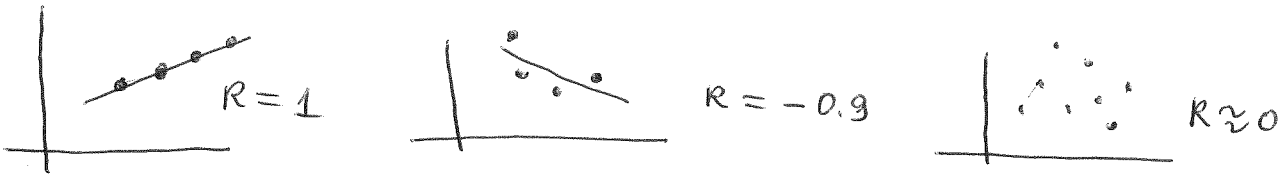
$$R^2 = \frac{Cov^2(x,y)}{VAR\,x \cdot VAR\,y}$$

במקרה שלנו $R^2 = \frac{2^2}{1.25 * 3.25} = 0.985$ מהיכן ש $VAR(y) = 3.25$

ניתן להוכיח שמקדם המתאם $-1 \le R \le 1$

הסימן של R הוא כמו הסימן של <u>השונות המשותפת</u>. סימן חיובי פרושו ש x ו-y באותה מגמה, בעוד שסימן שלילי מעיד על קשר שלילי בלומר 2 המשתנים במגמות מנגדות (קו יורד). ערכו המוחלט של R מצביע על מידת ההתאמה של הנקודות לקו. אם $R = \mp 1$ זה מעיד שהנקודות כולן על הקו. ערכים מוחלטים

קרובים ל 1 מצביעים על קשר חזק. ערכים של R קרובים לאפס מצביעים על חוסר קשר וערכים אקראיים.



* ניתן לפתור את הרגרסיה בחלוקות רבעוניות ופרוונות. הרבעון הוא רבעון מס 2 והשאלות נפתרו באמצעות המשוואות הקנוניות של הרגרסיה. בדומה לפתרון באמצעות המשוואות הסטטיסטיות ולקבל תוצאות זהות.

רבעון מס 3. על פי נתוני רבעון מס 2 נקבל הקו $y=1.5+1.6x$

מתוך אותה טבלת נתונים מצא את משוואת הקו $x=\alpha_0+\alpha_1 y$

מתקיימים 2 התנאים מן הגזירה $\bar{x}=\alpha_0+\alpha_1\bar{y} \Leftarrow 2.5=\alpha_0+5.5\alpha_1$

$$\alpha_1=\frac{Cov(x,y)}{VAR(y)}=\frac{2}{VAR\,y}=\frac{2}{3.25}=0.615$$

$$\alpha_0=2.5-0.615*5.5=-0.8825$$

באשר $\frac{\sum y_i^2}{n}-\bar{y}^2=VAR\,y=3.25$

בלומר משוואת הקו תהיה $\hat{x}=-0.8825+0.615y$

א. מתי ניתן להגיע את x ממשוואת הקו $y=\beta_0+\beta_1 x$? רק כאשר כל הנקודות הן על הקו בלומר $R^2=1$

ב. בדומה לבדר ש R^2 הוא מכפלת שפועי 2 הקווים $y(x)$, $x(y)$ שכן

$$R^2=\frac{Cov^2(x,y)}{VAR\,x\cdot VAR\,y}=\frac{Cov(x,y)}{VAR\,x}\cdot\frac{Cov(x,y)}{VAR\,y}=\beta_1\cdot\alpha_1$$

ג. מתוך משוואת הקווים שנתקבלו מאותה טבלת נתונים $y=1.5+1.6x$ והקו $x=-0.8825+0.615y$ קבע לאי בדור גבוה יותר למשתנה x או למשתנה y

$$1.6=\frac{Cov(x,y)}{VAR\,x}\quad,\quad 0.615=\frac{Cov(x,y)}{VAR\,y}$$

מכאן מתקבל $VAR\,y>VAR\,x$ בלומר הבדור של y גבוה יותר