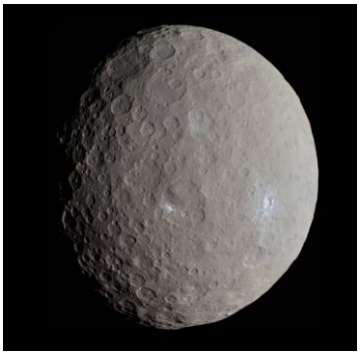


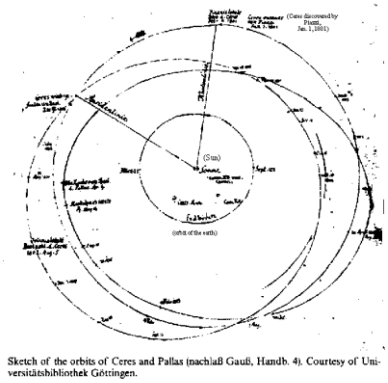

מתאם ורגרסיה לינארית

כוכב חדש



צרס – כפי שצולם ב-2016

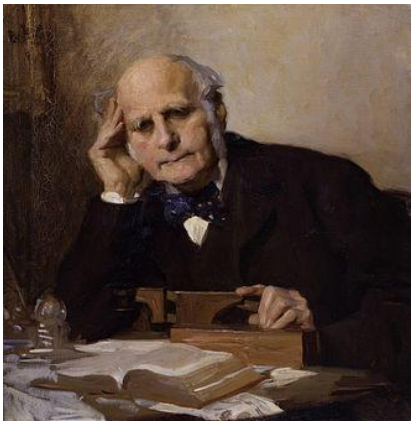
קרל פרידריך גאוס (1855-1877)



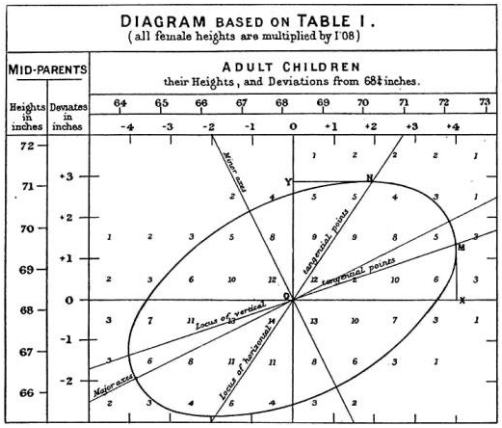
שרטוט מסלול צרס סביב השמש (גאוס)

Sketch of the orbits of Ceres and Pallas (nachlaß Gauß, Handb. 4). Courtesy of Universitätsbibliothek Göttingen.

הקשר בין גובה אבות וגובה בניהם – רגרסיה לממוצע



סיר פרנסיס גאלטון (1822-1911)



מתאם בין גבהים של אבות ובנים - 1886

רגרסיה לממוצע

דניאל כהנמן
חתן פרס נובל לכלכלה - 2002

> JAMA. 2005 Jul 13;294(2):218-28. doi: 10.1001/jama.294.2.218.

Contradicted and initially stronger effects in highly cited clinical research

John P A Ioannidis [1]

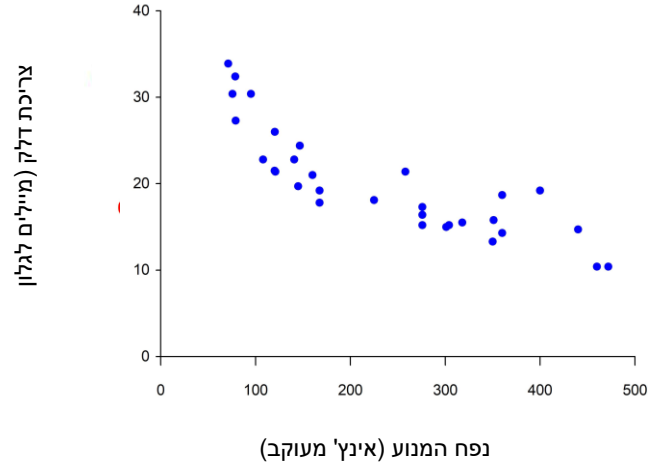
Affiliations + expand

PMID: 16014596 DOI: 10.1001/jama.294.2.218

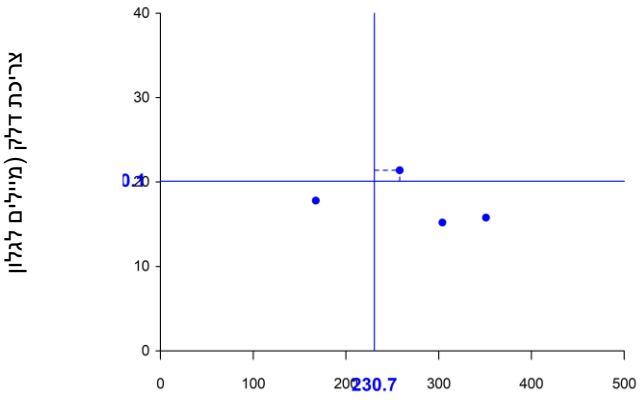
Abstract

Context: Controversy and uncertainty ensue when the results of clinical research on the effectivenes of interventions are subsequently contradicted. Controversies are most prominent when high-impac research is involved.

דוגמא נתוני המכוניות



דוגמא: נתוני המכוניות ביחס לממוצעים



נתוני המכוניות - חישובים

	mpg	disp	הפרש מהממוצע mpg	הפרש מהממוצע disp	הפרש מהממוצע mpg	הפרש מהממוצע disp	מכפלה
Cadillac Fleetwood	10.4	472	10.4-20.1	472-230.7	-9.7	241.3	-2340.61
Hornet 4 Drive	21.4	258	21.4-20.1	258-230.7	1.3	27.3	35.49
Merc 280C	17.8	167.6	17.8-20.1	167.6-230.7	-2.3	-63.1	145.13
Datsun 710	22.8	108	22.8-20.1	108-230.7	2.7	-122.7	-331.29

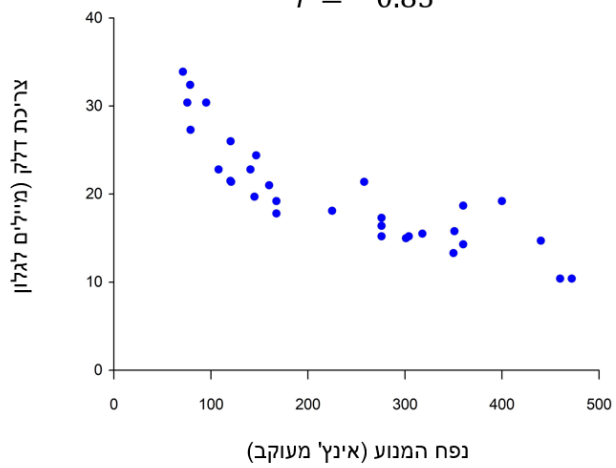
$$\text{cov}(mpg, disp) = \frac{(10.4-20.1)\cdot(472-230.7)+(21.4-20.1)\cdot(258-230.7)+\cdots.}{31} = \frac{-19626.02}{31} = -633.1$$

$$s_{mpg} = 6.03 \quad s_{disp} = 123.9$$

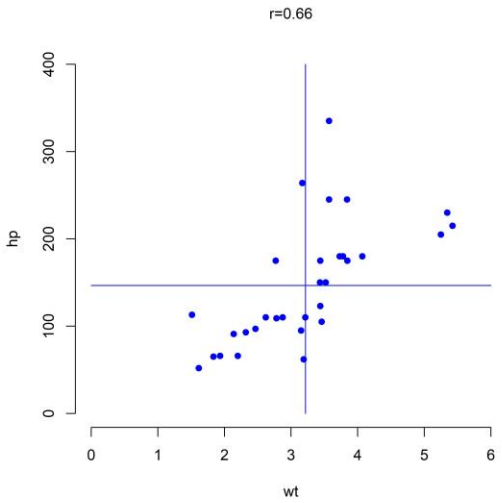
$$r(mpg, disp) = r = \frac{\text{cov}(mpg, disp)}{s_{mpg}\cdot s_{disp}} = \frac{-19626.02}{6.03\cdot 123.9} = -0.85$$

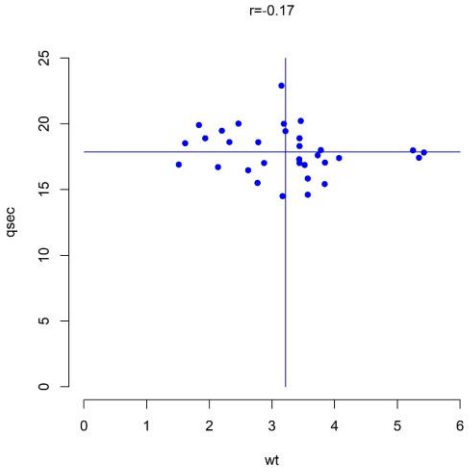
סיכום עד כה

$$r = -0.85$$



נתוני המכוניות: משתנים נוספים

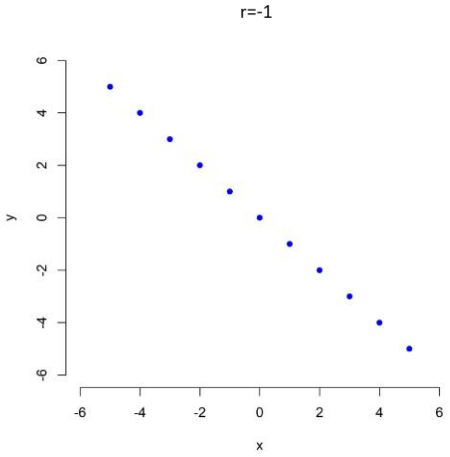


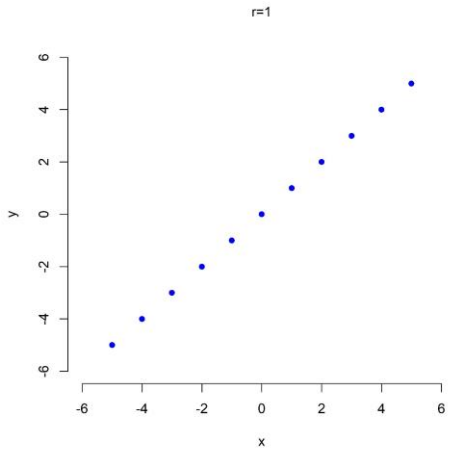


נתוני המכוניות: כל המתאמים

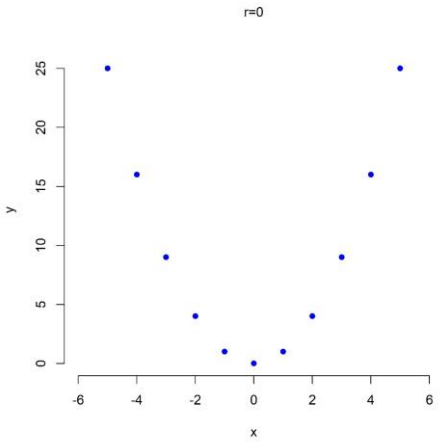
	mpg	disp	hp	drat	wt	qsec
mpg	1.00	-0.85	-0.78	0.68	-0.87	0.42
disp	-0.85	1.00	0.79	-0.71	0.89	-0.43
hp	-0.78	0.79	1.00	-0.45	0.66	-0.71
drat	0.68	-0.71	-0.45	1.00	-0.71	0.09
wt	-0.87	0.89	0.66	-0.71	1.00	-0.17
qsec	0.42	-0.43	-0.71	0.09	-0.17	1.00

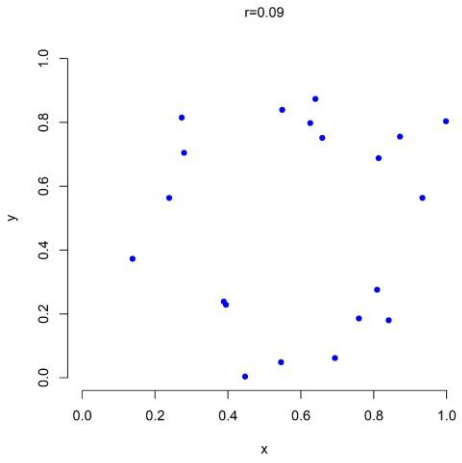
קשר לינארי מלא





מתאם אפס

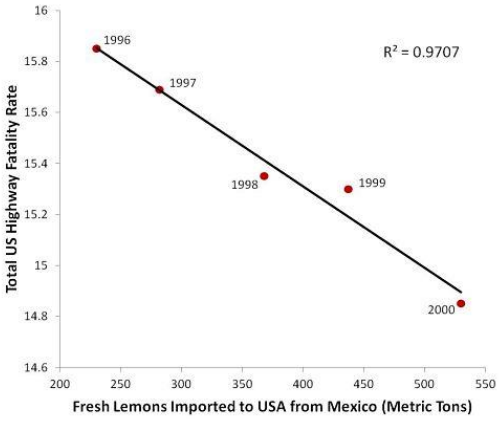




תכונות מקדם המתאם – "מתאם פירסון"

- מקדם המתאם ניתן לחישוב רק עבור משתנים בסולם רווח או מנה
- מקדם המתאם מודד את עוצמת הקשר הלינארי בין שני משתנים
- מקדם המתאם הוא סימטרי
- ערכו של מקדם המתאם הינו תמיד בין 1- ל-1
- אם קיים קשר לינארי חיובי מלא בין המשתנים אז מקדם המתאם שווה ל-1
- אם קיים קשר לינארי שלילי מלא בין המשתנים אז מקדם המתאם שווה ל-1-
- אם מקדם המתאם שווה ל-0, עדיין ייתכן כי קיים קשר לא לינארי בין המשתנים
- ערכים שבין 1- ל-1 נתונים לפרשנות

מתאם אינו מספיק כדי לקבוע סיבתיות!



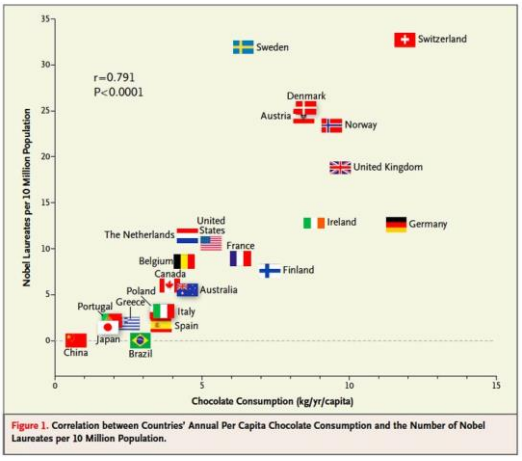
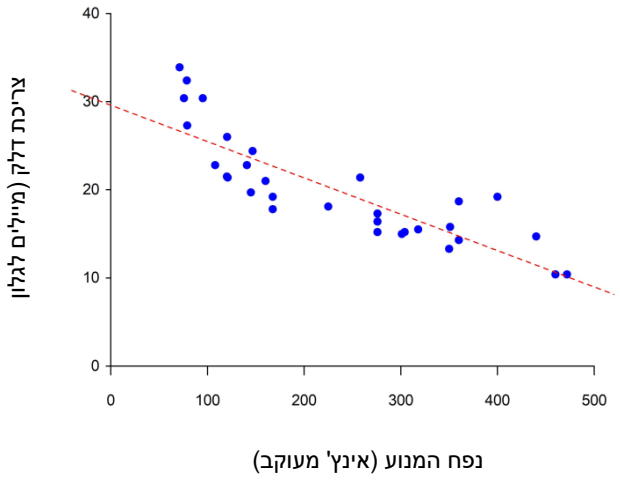
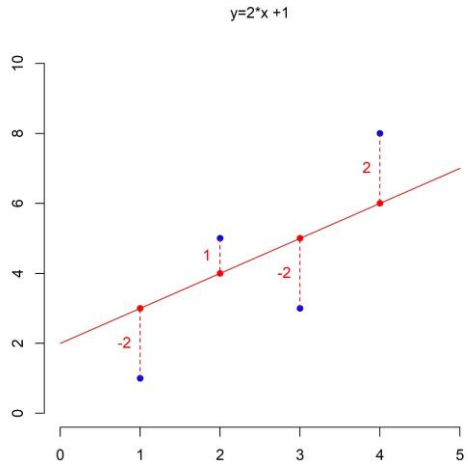


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

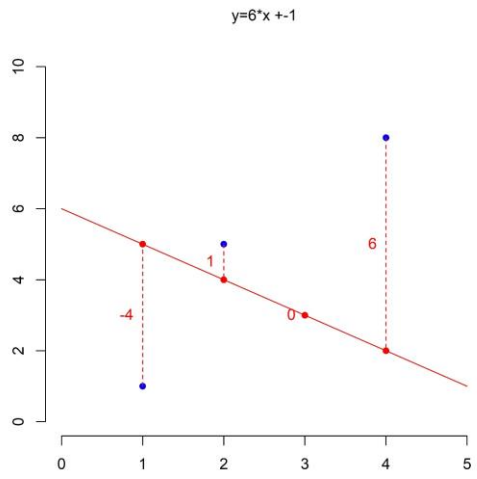
בחזרה לנתוני המכוניות



דוגמה מלאכותית



המשך דוגמה מלאבותית



אמדן הריבועים הפחותים

- הרעיון: להביא למינימום את סכום ריבועי ההפרשים
- בדוגמה המלאבותית שלנו: ערכי x הם 1,2,3,4 וערכי y המתאימים הם 1,5,3,6
- נחפש ערכים a ו-b כך שהסכום הבא יהיה מינימלי:

$$\text{Sum of Squares} = [1-(a\cdot 1+b)]^2 + [5-(a\cdot 2+b)]^2 + [3-(a\cdot 3+b)]^2 + [8-(a\cdot 4+b)]^2$$

קצת תיאוריה

$$Y = a + b \cdot X + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

- X ו-Y הם משתנים כמותיים
- קיים קשר לינארי בין X ל-Y
- X נמדד ללא טעויות מדידה
- Y נמדד עם טעות מדידה
- ההתפלגות של טעות המדידה היא נורמלית עם תוחלת אפס
- השונות של טעות המדידה אינה תלויה ב-Y

אמדני הריבועים הפחותים

אמדני הריבועים הפחותים:

$$\hat{b} = \frac{\mathrm{cov}(x,y)}{s_x^2} \qquad \hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$$

הקשר בין הרגרסיה ומקדם המתאם:

$$\hat{b} = \frac{\mathrm{cov}(x,y)}{s_x} \cdot \frac{1}{s_x} \cdot \frac{s_y}{s_y} = \frac{\mathrm{cov}(x,y)}{s_x \cdot s_y} \cdot \frac{s_y}{s_x} = \frac{s_y}{s_x} \cdot r_{xy}$$

בדוגמת המכוניות

$\overline{disp} = 230.72 \quad \overline{mpg} = 20.09 \qquad s_{disp} = 123.9 \quad s_{mpg} = 6.03$

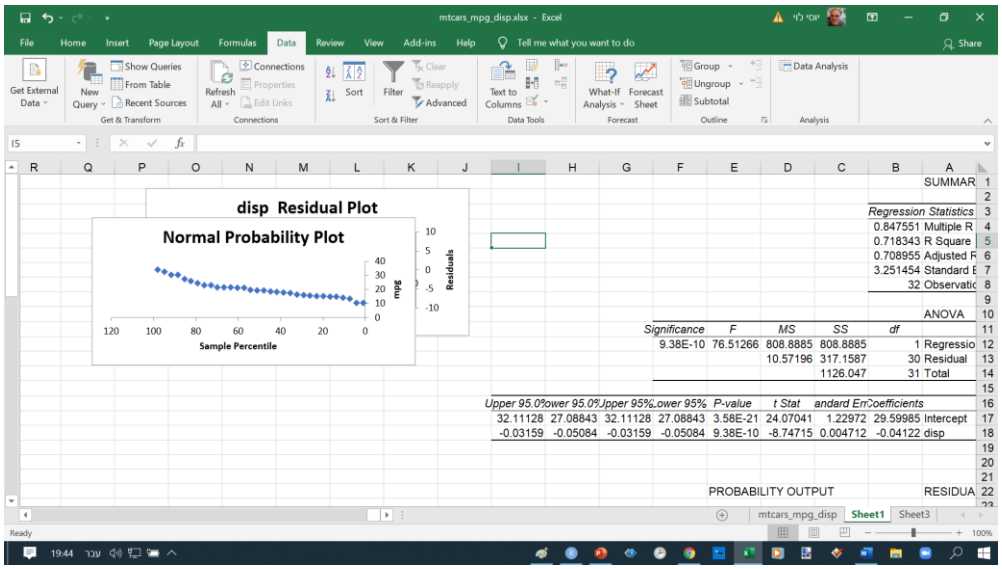
$cov(mpg, disp) = -633.1 \qquad r = -0.85$

$\hat{b} = \frac{-633.1}{123.9^2} = -0.0412 \qquad \hat{b} = \frac{6.03}{123.9} \cdot (-0.85) = -0.0412$

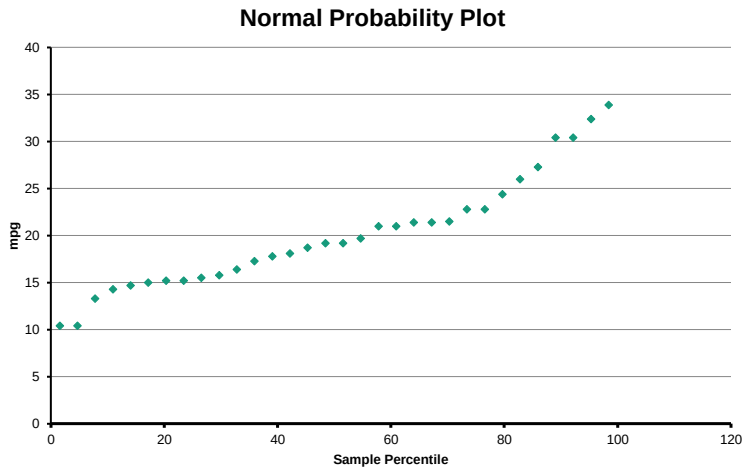
$\hat{a} = 20.09 + 0.0412 \cdot 230.72 = 29.599$

$$mpg = 29.599 - 0.0412 \cdot disp$$

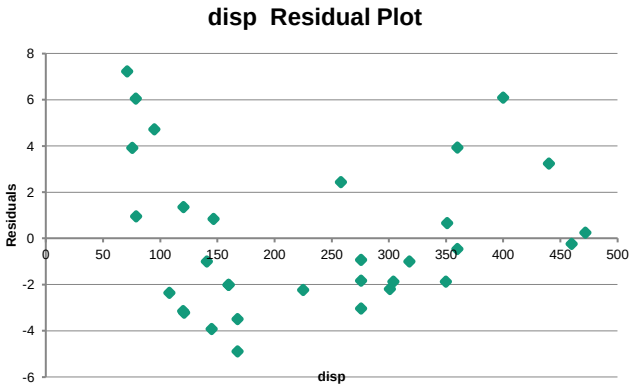
פלט אקסל

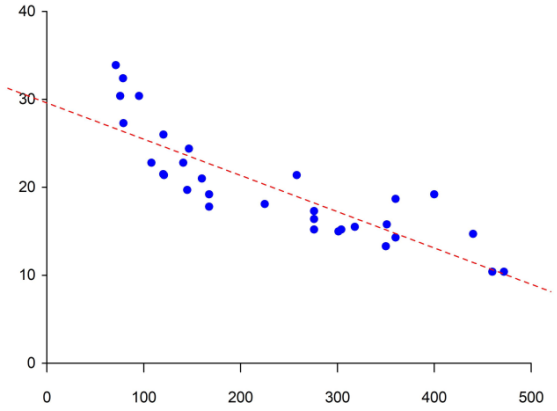


תרשים התפלגות נורמלית

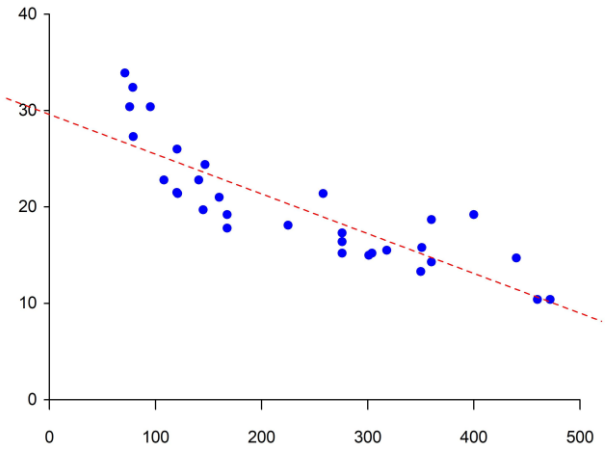


דיאגרמת שאריות





סיכום קצר



לוח ניתוח שונות

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	808.888	808.888	76.5127	9.38033E-10
Residual	30	317.159	10.572		
Total	31	1126.05			

$$\frac{1126.05}{31} = 36.324 = \mathrm{V(mpg)}$$

סכום ריבועי השאריות

$$\frac{808.888}{1126.05} = 0.7183 = (-0.847)^2 = r^2$$

אחוז השונות המוסברת:

אמידת סטיית התקן של השאריות

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	808.888	808.888	76.5127	9.38033E-10
Residual	30	317.159	10.572		
Total	31	1126.05			

דרגות החופש

סכום ריבועי השאריות

דרגות החופש= מספר התצפיות – מספר הפרמטרים הנאמדים df=32-2=30

MSE=סכום ריבועי השאריות/דרגות החופש $\text{Standard error} = \sqrt{\text{MSE}} = \sqrt{10.572} = 3.2515$

אחוז השונות המוסברת: $808.888/1126.05 = 0.7183$

מדדים שונים

Regression Statistics	
Multiple R	0.8476
R Square	0.7183
Adjusted R Square	0.709
Standard Error	3.2515
Observations	32

מקדם המתאם:

אחוז השונות המוסברת:

סטיית התקן של השאריות

אמדני הריבועים הפחותים – "מקדמי הרגרסיה"

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	$\hat{a}=$ 29.600	$SD(\hat{a})$ 1.230	24.070	0.0000	27.088	32.111
disp	$\hat{b}=$ -0.041	$SD(\hat{b})$ 0.005	-8.747	0.0000	-0.051	-0.032

השערות דו צדדיות
$H_0: a = 0$ $H_0: b = 0$

רווחי סמך עבור a ו-b

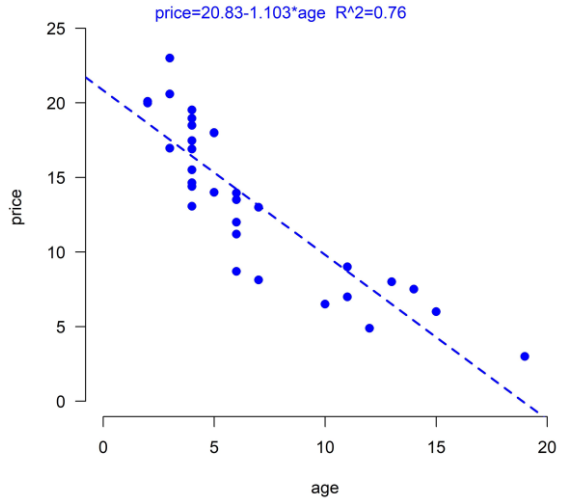
משוואת הרגרסיה: mpg=29.6-0.041·disp

מה המשמעות של הפרמטרים?

דוגמה: מחירי מכוניות

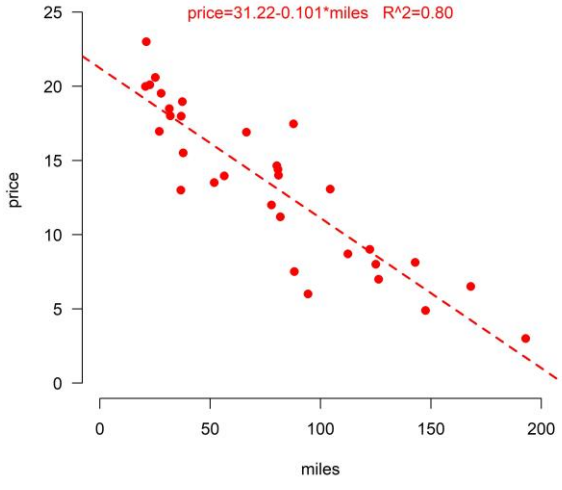
year	price	age	miles
2010	6.99	11	126.331
2017	16.9	4	66.462
2018	20.588	3	25.152
2015	11.198	6	81.792
2017	19.516	4	27.825
2010	8.995	11	122.354
2016	17.995	5	31.953
2017	14.385	4	80.756
2002	2.995	19	193
2018	16.957	3	26.95
2014	12.99	7	36.705
2006	5.995	15	94.397
2009	4.888	12	147.598
2017	14.633	4	80.209
2008	7.998	13	125.065
2015	13.95	6	56.409

מחיר על פי גיל הרכב



31

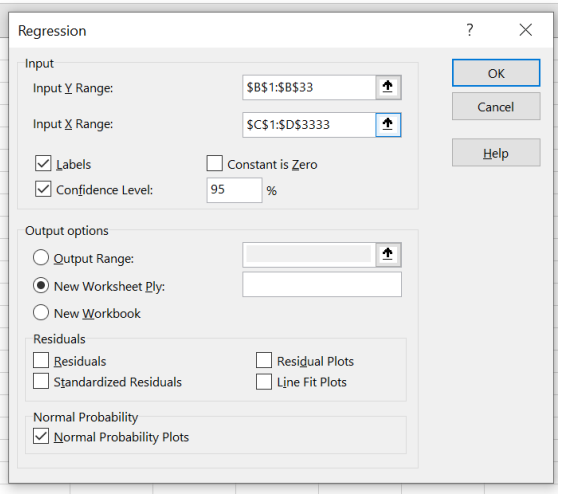
מחיר על פי המרחק שהמכונית עברה



32

רגרסיה מרובה: מחיר על פי הגיל ומרחק הנסיעה

	A	B	C	D
	year	price	age	miles
	2010	6990	11	126331
	2017	16900	4	66462
	2018	20588	3	25152
	2015	11198	6	81792
	2017	19516	4	27825
	2010	8995	11	122354
	2016	17995	5	31953
	2017	14385	4	80756
	2002	2995	19	193000
	2018	16957	3	26950
	2014	12990	7	36705
	2006	5995	15	94397
	2009	4888	12	147598
	2017	14633	4	80209
	2008	7998	13	125065
	2015	13950	6	56409
	2017	18955	4	37450
	2017	17464	4	87764
	2015	8699	6	112460
	2018	22995	3	21053
	2017	13064	4	104397
	2016	17981	5	36882



רגרסיה מרובה: מחיר על פי הגיל ומרחק הנסיעה

Regression Statistics	
Multiple R	0.943
R Square	0.888
Adjusted R Square	0.881
Standard Error	1.842
Observations	32

אחוז השונות המוסברת: $783.11/881.47 = 0.888$

סטיית התקן של השאריות $\sqrt{\text{MSE}} = \sqrt{3.39} = 1.842$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	783.11	391.55	115.44	0.0000
Residual	29	98.36	3.39		
Total	31	881.47			

דרגות החופש= מספר התצפיות – מספר הפרמטרים הנאמדים: df=32-3=29

MSE=סכום ריבועי השאריות/דרגות החופש: $98.36/29 = 3.39$

רגרסיה מרובה: מחיר על פי הגיל ומרחק הנסיעה

	Coefficients	Std. Err	t Stat	P-value	Lower 95%	Upper 95%
Intercept	22.077	0.652	33.861	0.0000	20.743	23.410
age	-0.579	0.120	-4.828	0.0000	-0.825	-0.334
miles	-0.062	0.011	-5.768	0.0000	-0.084	-0.040

משוואת הרגרסיה: $price = 22.077 - 0.579 \cdot age - 0.062 \cdot miles$

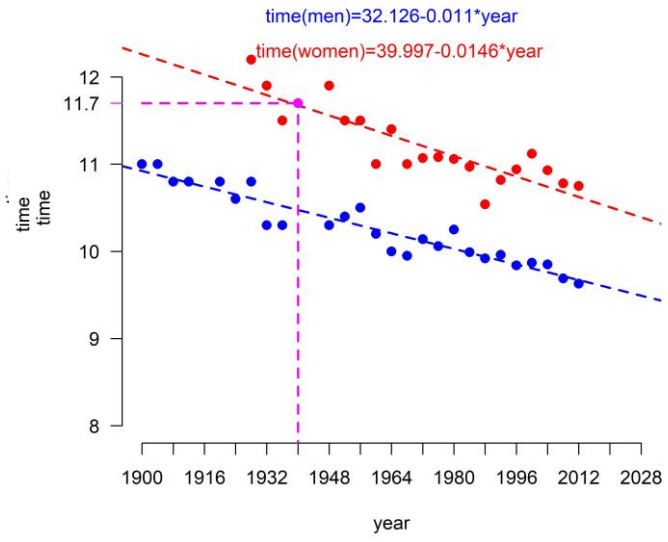
יש מתאם בין המרחק שהמכונית עברה ובין הגיל שלה. האם שני המשתנים הכרחיים?

דוגמה: זמני המנצחים בריצת 100 מטר במשחקים האולימפיים

ג'סי אוונס – ברלין 1936

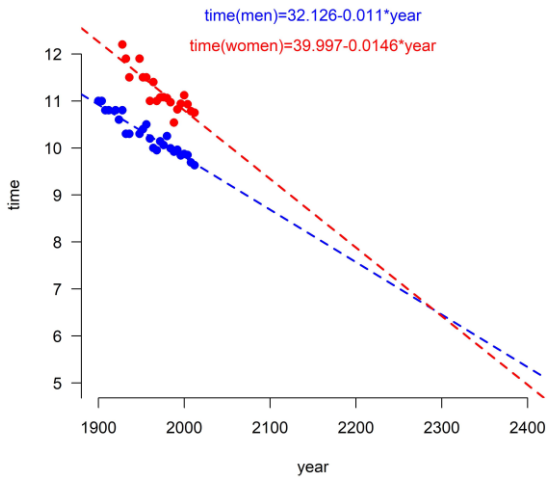
city	year	men	women
Paris	1900	11	
St. Louis	1904	11	
London	1908	10.8	
Stockholm	1912	10.8	
Antwerp	1920	10.8	
Paris	1924	10.6	
Amsterdar	1928	10.8	12.2
Los Angel	1932	10.3	11.9
Berlin	1936	10.3	11.5
London	1948	10.3	11.9
Helsinki	1952	10.4	11.5
Melbourne	1956	10.5	11.5
Rome	1960	10.2	11
Tokyo	1964	10	11.4
Mexico Cit	1968	9.95	11

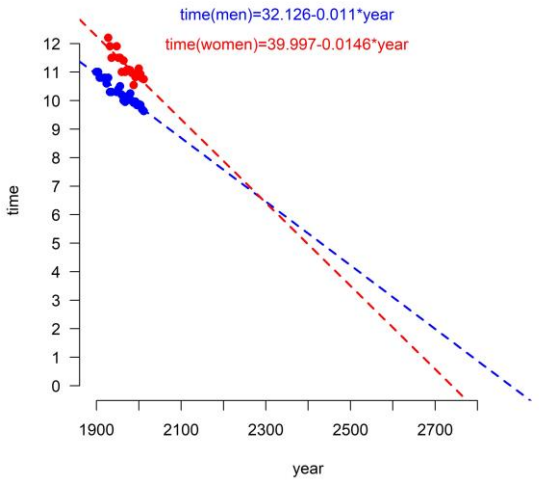
זמני המנצחים בריצת 100 מטר במשחקים האולימפיים



37

חיזויים מסוכנים





38

זמני ריצה – גישה אחרת

- הנחה: קצב השיפור של זמני הריצה שווה לשני המינים
- משמעות: קווי הרגרסיה של שני המינים מקבילים
- במשואות הרגרסיה, הערך של b לשני המינים, טבל לכל מין יש ערך שונה של a:

$$time(man) = a_1 + b \cdot year$$

$$time(woman) = a_2 + b \cdot year$$

- משמעות שניה: ההבדל בין השנים והגברים לא משתנה לאורך השנים

זמני ריצה – קווי רגרסיה מקבילים

year	women	time
1900	0	11
1904	0	11
1908	0	10.8
1912	0	10.8
1920	0	10.8
1924	0	10.6
1928	0	10.8
1928	1	12.2
1932	0	10.3
1932	1	11.9
1936	0	10.3
1936	1	11.5
1948	0	10.3
1948	1	11.9
1952	0	10.4
1952	1	11.5
1956	0	10.5
1956	1	11.5
1960	0	10.2

- מגדירים משתנה חדש בשם woman
- עבור נשים ערכו של משתנה זה שווה ל-1
- עבור גברים ערכו של משתנה זה שווה לאפס
- כעת אפשר לשים את כל הזמנים בעמודה אחת
- כעת אפשר לשים את כל הזמנים בעמודה אחת

קווי רגרסיה מקבילים – משוואות הרגרסיה

- משוואת הרגרסיה תהיה

$$time = a + b_1 \cdot woman + b_2 \cdot year$$

- עבור גברים, ערכו של המשתנה woman תמיד שווה ל-0, ולכן עבור גברים משוואת הרגרסיה היא

$$time(man) = a + b_1 \cdot 0 + b_2 \cdot year = a + b_2 \cdot year$$

- עבור נשים, ערכו של המשתנה woman תמיד שווה ל-1, ולכן עבור נשים משוואת הרגרסיה היא

$$time(women) = a + b_1 \cdot 1 + b_2 \cdot year = (a + b_1) + b_2 \cdot year$$

קווי רגרסיה מקבילים – משוואות הרגרסיה

	Coefficients	Std. Err	t Stat	P-value	Lower 95%	Upper 95%
Intercept	$\hat{a} =$ 34.998	1.935	18.084	0.0000	31.081	38.916
women	$\hat{b}_1 =$ 1.076	0.061	17.658	0.0000	0.953	1.200
year	$\hat{b}_2 =$ -0.013	0.001	-12.781	0.0000	-0.015	-0.011

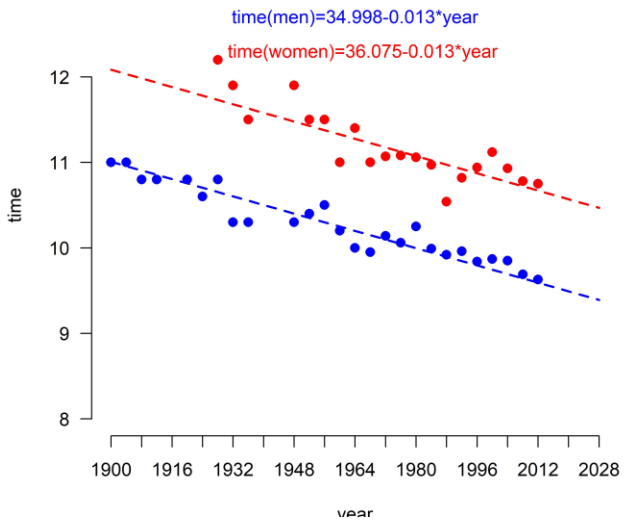
- עבור גברים משוואת הרגרסיה היא

$$time(man) = a + b_2 \cdot year = 34.998 - 0.013 \cdot year$$

- עבור נשים משוואת הרגרסיה היא

$$time(women) = (a + b_1) + b_2 \cdot year = 36.075 - 0.013 \cdot year$$

קווי רגרסיה מקבילים



מודל יותר מציאותי