רגרסיה ליניארית מרובת משתנים

* מבוססת על הרצאתו של פרופ' הלל בר-גרא, אונ' בן גוריון בנגב

דוגמה לבעית רגרסיה פשוטה

נאספו נתונים על התפתחות הלב של ילד מגיל 6 ועד 17. בערך כל חצי שנה נערכה מדידה פלורוסקופית של הקוטר הרוחבי של הלב בס"מ, סה"כ 22 תצפיות.

הנתונים מופיעים בטבלה הבאה כאשר:

עמודה 1: גיל בחודשים

עמודה 2: לוגריתם הקוטר הרוחבי של הלב בס"מ

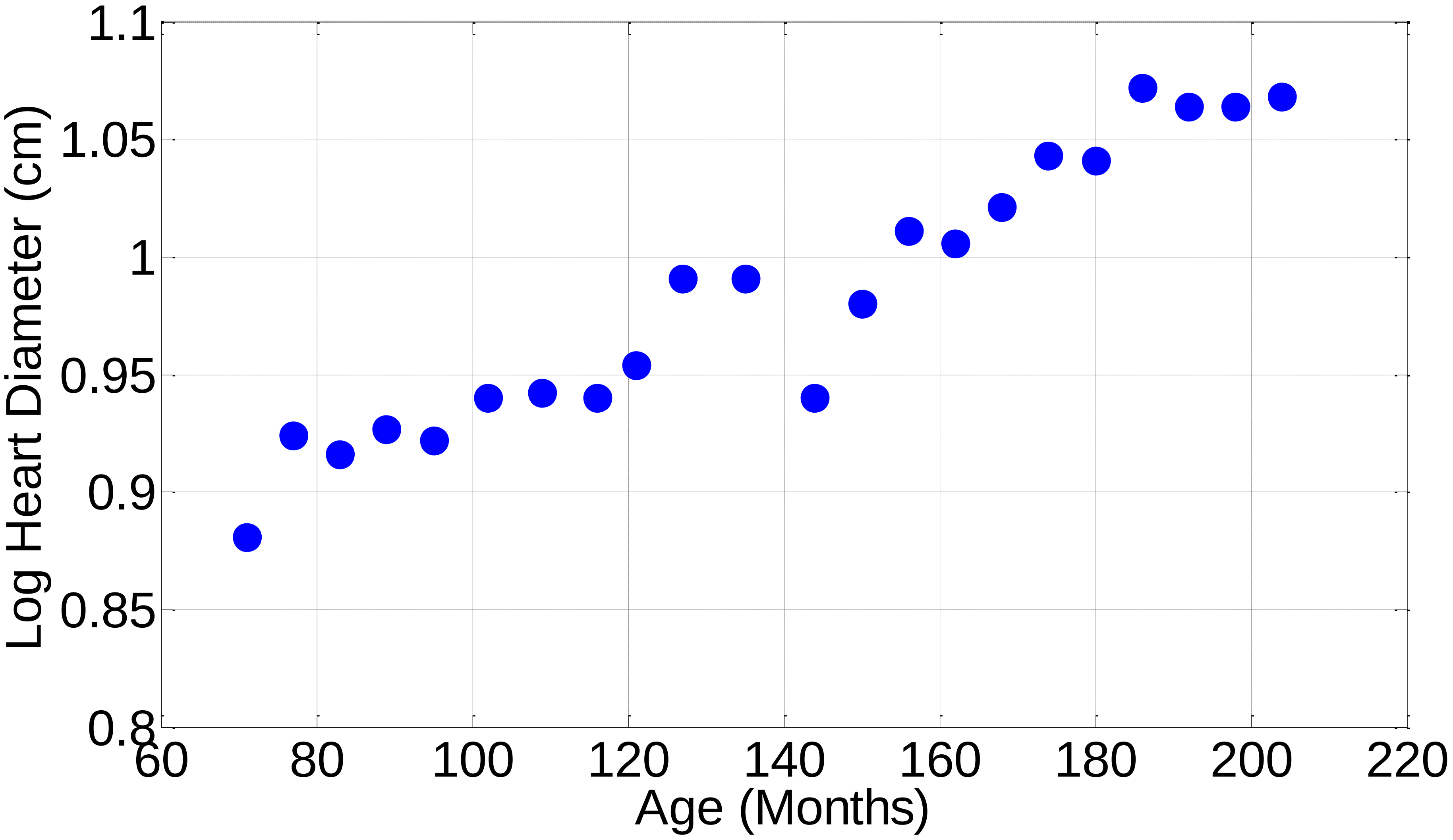
נתוני גיל וקוטר הלב

גיל	קוטר
71	0.881
77	0.924
83	0.916
89	0.927
95	0.922
102	0.94
109	0.942
116	0.94

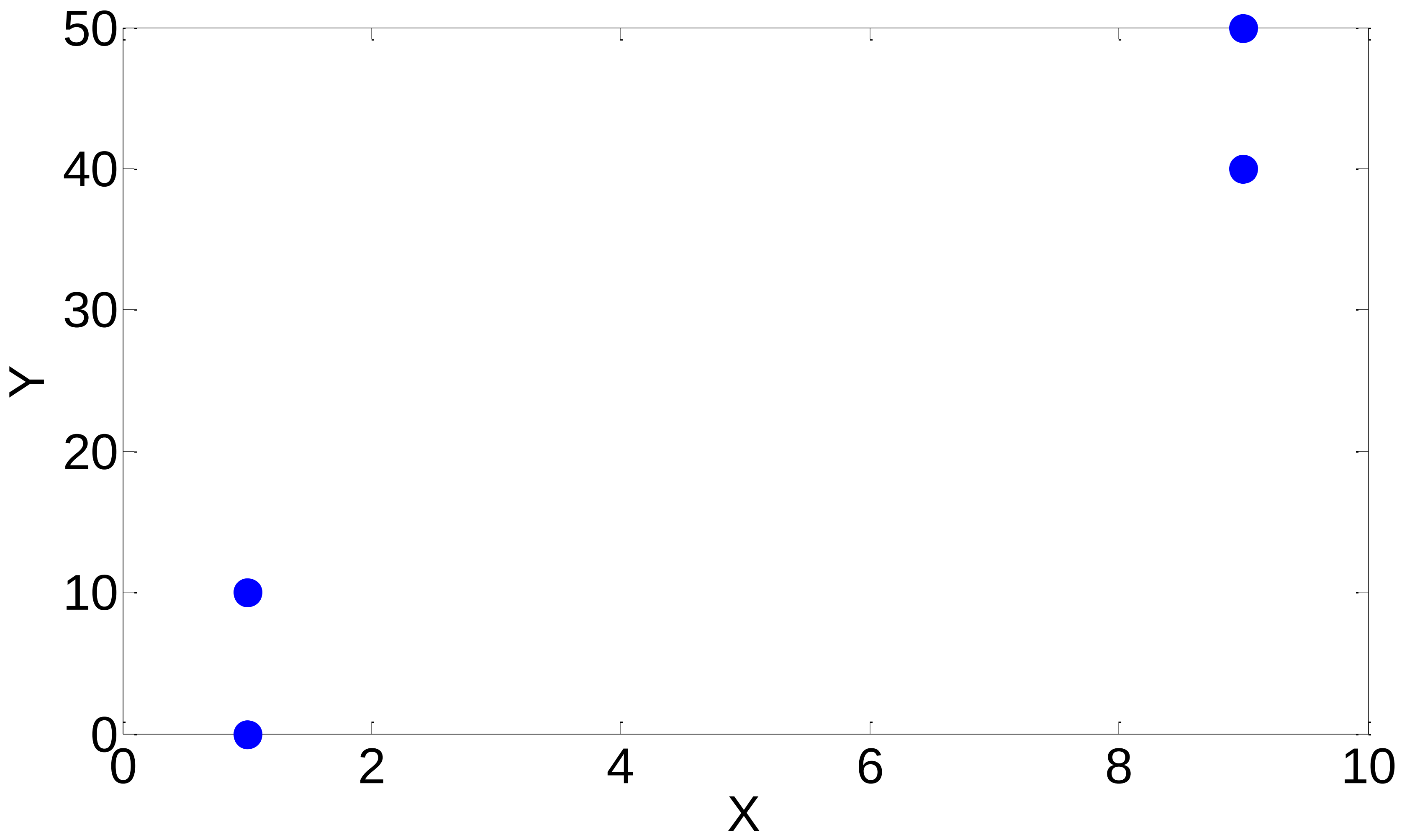
גיל	קוטר
121	0.954
127	0.991
135	0.991
144	0.94
150	0.98
156	1.011
162	1.006
168	1.021

גיל	קוטר
174	1.043
180	1.041
186	1.072
192	1.064
198	1.064
204	1.068

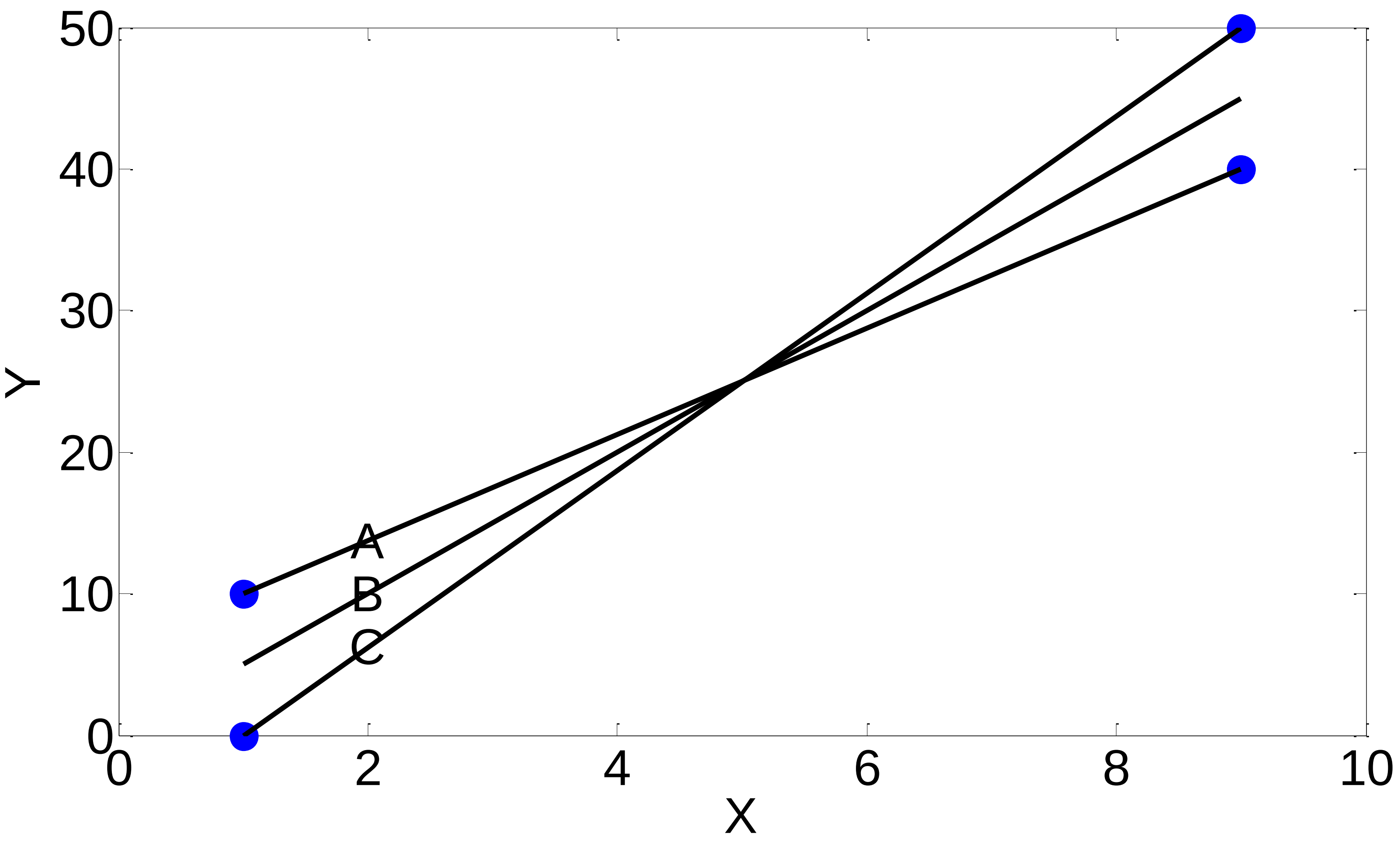
תרשים פיזור - קוטר לב לעומת גיל



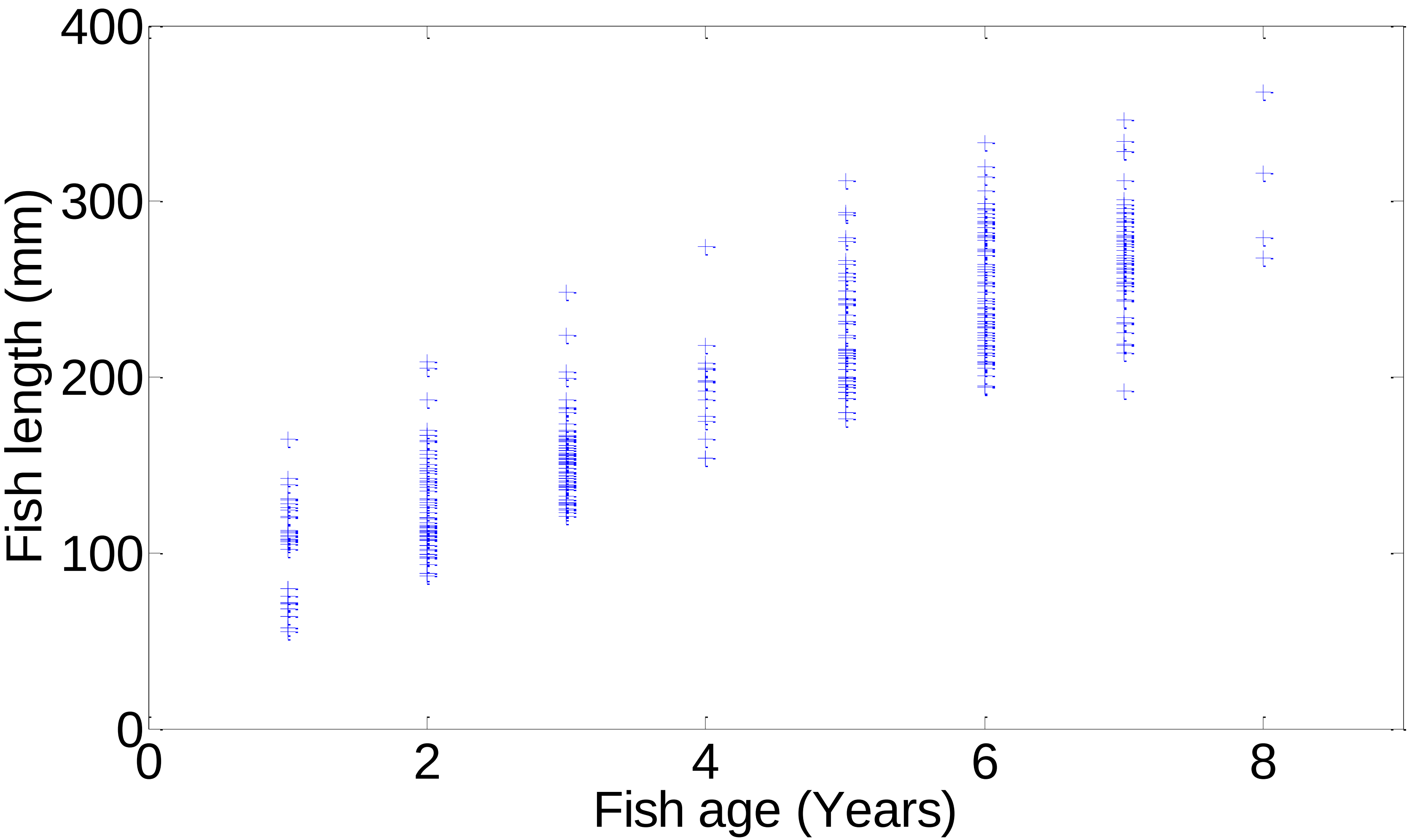
איך מעבירים קו ישר בין ארבע נקודות?



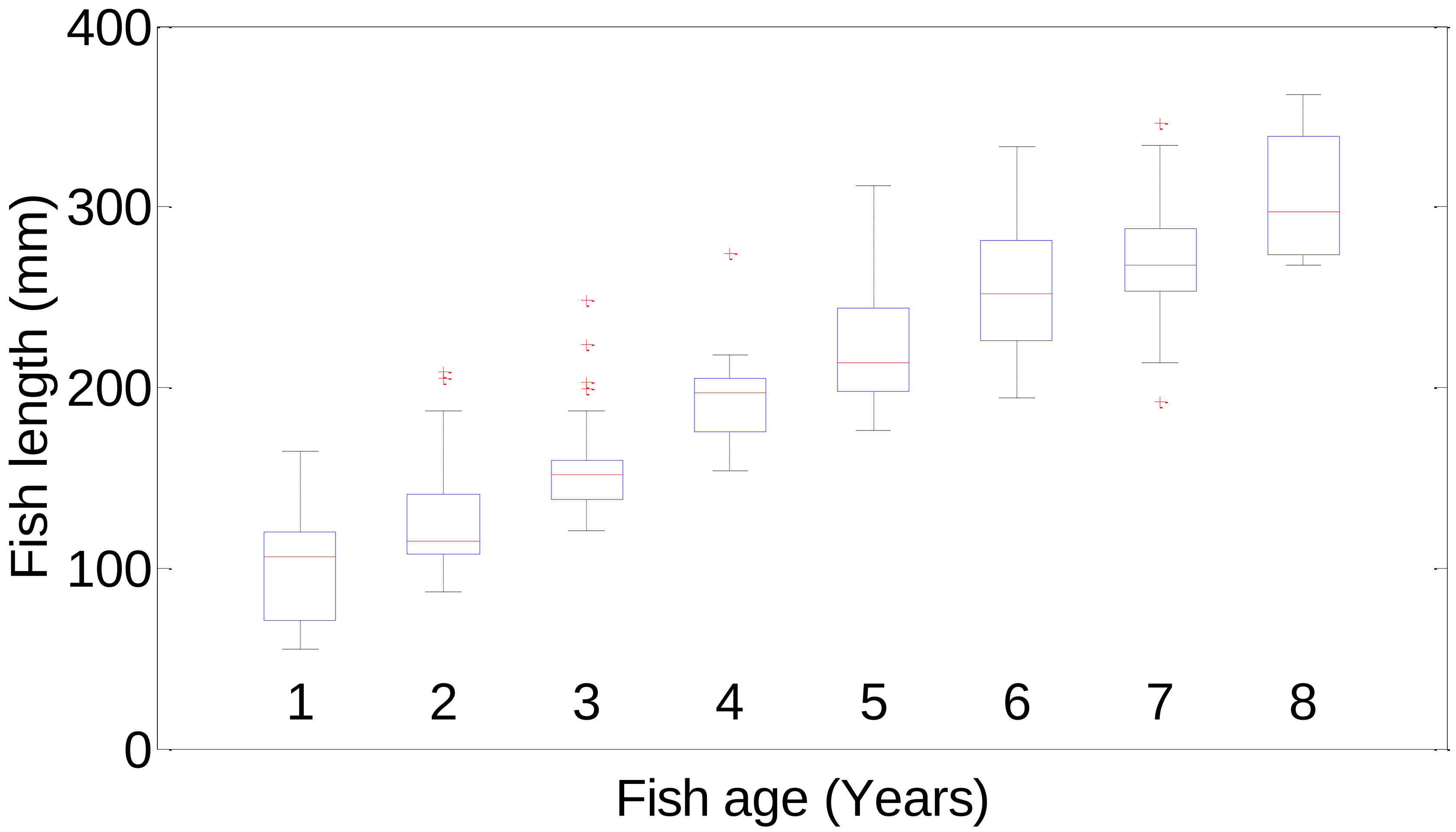
איך מעבירים קו ישר בין ארבע נקודות?



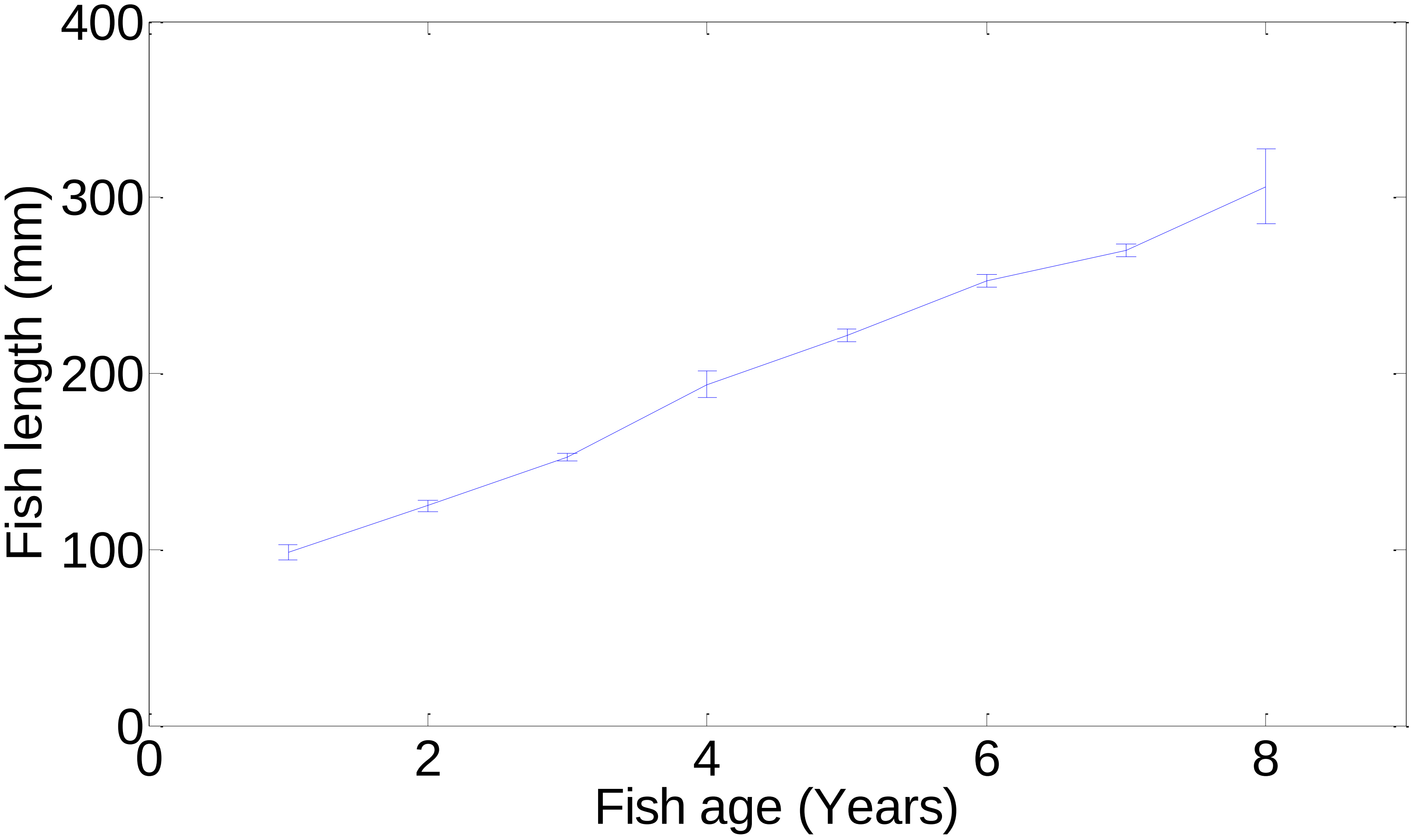
תהליך הגדילה של דגי בס



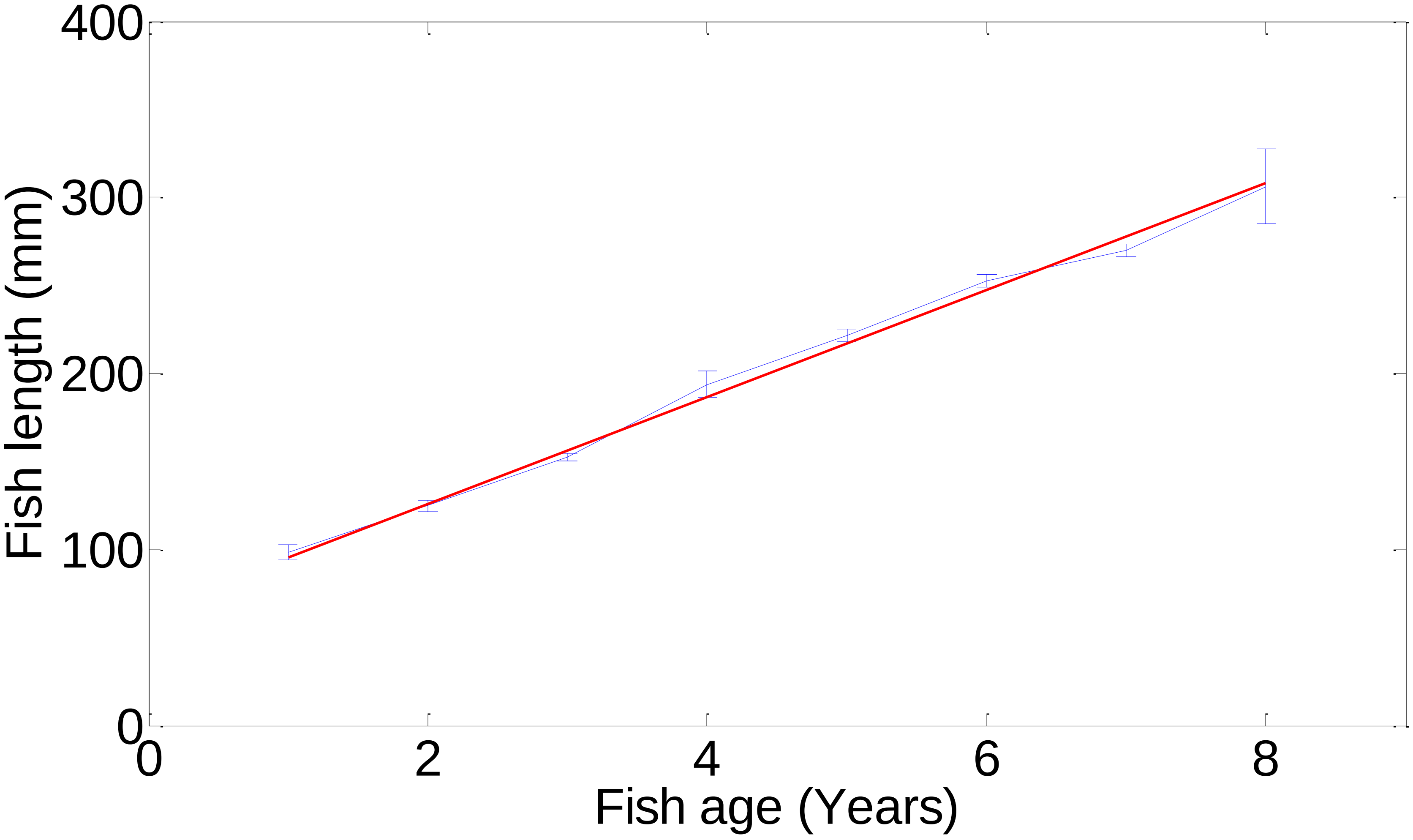
תרשים קופסא



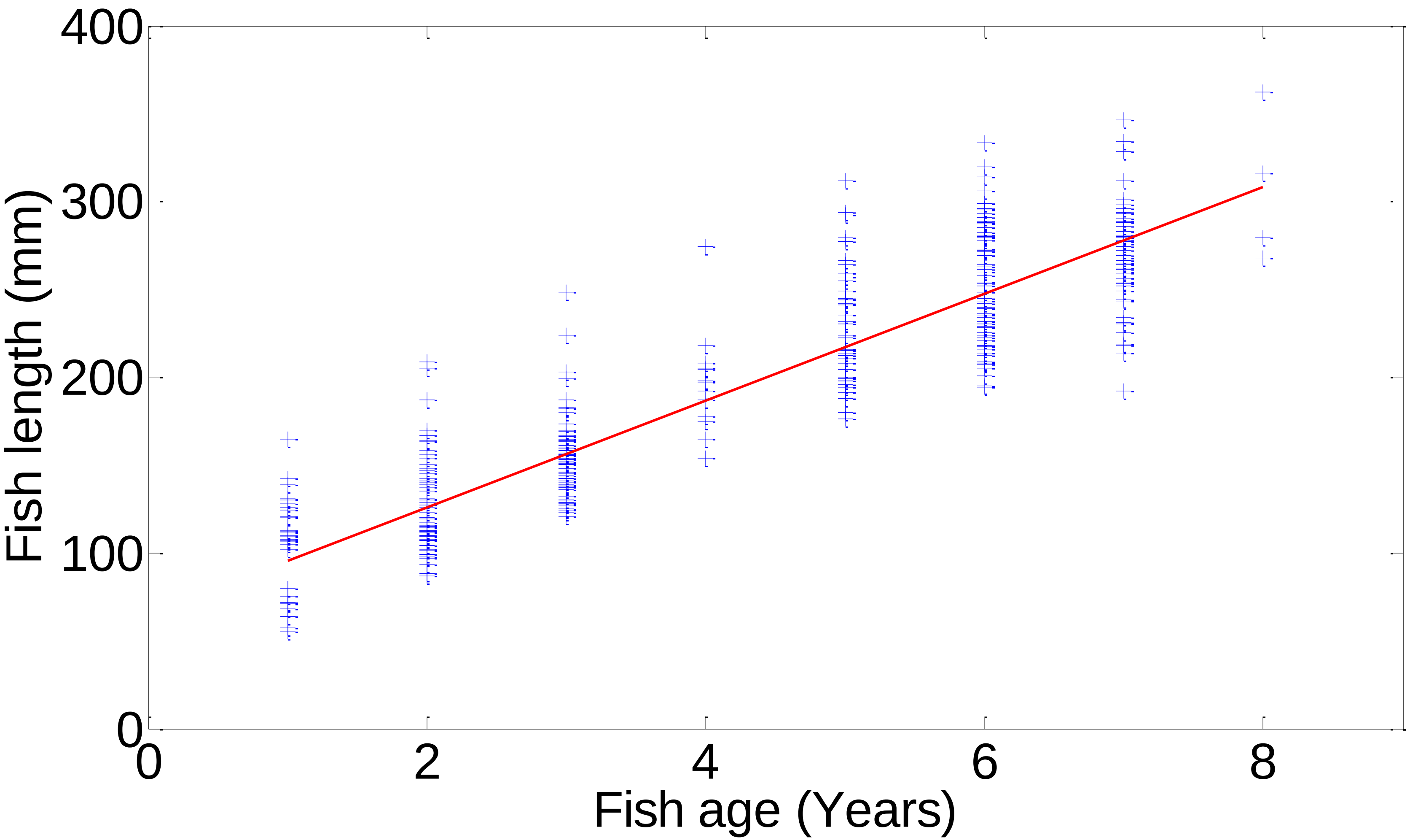
פונקצית תוחלת מותנה



פונקצית תוחלת מותנה ליניארית



קו הרגרסיה והנתונים המקוריים



מודל הרגרסיה הפשוטה

את מודל הרגרסיה הפשוטה ניתן לפרט ולכתוב בצורה הבאה:

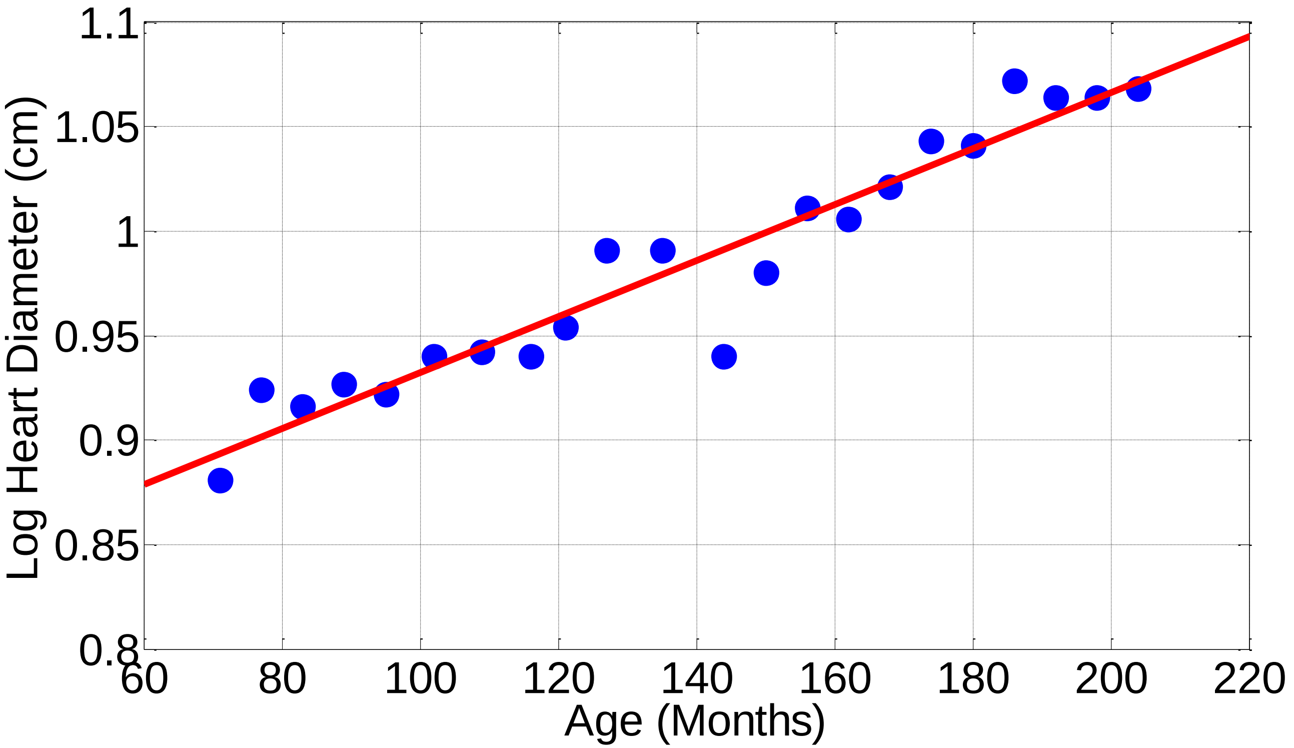
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_{i,} \quad \forall \ i = 1, \ldots, n \quad \varepsilon_i \sim N(0, \sigma^2)$

$$
\begin{array}{ccc}
y_1 & = & \beta_0 + \beta_1 x_1 + \varepsilon_1 \\
y_2 & = & \beta_0 + \beta_1 x_2 + \varepsilon_2 \\
\vdots & \vdots & \vdots \\
y_n & = & \beta_0 + \beta_1 x_n + \varepsilon_n
\end{array}
$$

קריטריון הריבועים הפחותים

הקריטריון המקובל להשוואה בין קווי רגרסיה הוא סכום ריבועי השגיאות:

$$Z(\beta_0,\beta_1)=\sum(y_i-\hat{y})^2=\sum[y_i-(\beta_0+\beta_1 x_i)]^2$$



קריטריון הריבועים הפחותים

הקריטריון המקובל להשוואה בין קווי רגרסיה הוא סכום ריבועי השגיאות:

$$Z(\beta_0,\beta_1)=\sum(y_i-\hat{y})^2=\sum[y_i-(\beta_0+\beta_1 x_i)]^2$$

על פי קריטריון זה, נבחר אומדים לפרמטרים β_0,β_1 כך שסכום ריבועי השגיאות יהיה מינימאלי. במילים אחרות נמזער את Z לפי β_0,β_1. כדי למצוא את אמדי הריבועים הפחותים, נגזור את Z ונשווה לאפס:

$$\frac{\partial Z}{\partial \beta_0}=\quad -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)\quad =0$$

$$\frac{\partial Z}{\partial \beta_1}=\quad -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)x_i\quad =0$$

המשוואות הנורמליות

פתרון בעיית המזעור הינו האומדים $\hat{\beta}_0$ ו- $\hat{\beta}_1$

המקיימים את מערכת המשוואות הבאות:

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0\\ \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i &= 0\end{aligned}$$

משוואות אלה נקראות המשוואות הנורמליות.

פתרון המשוואות הנורמליות

נסמן:

$$SXX = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad SXY = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$S_X = \sqrt{S_{XX}} \qquad S_{XY} = SXY/(n-1)$$

כדאי לדעת:

$$SXY = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(y_i - \bar{y})x_i = \sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y}$$

$$SXX = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2$$

הפתרון המתקבל הוא:

$$\hat{\beta}_1 = \frac{SXY}{SXX} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

נתוני גיל וקוטר הלב

גיל	קוטר
71	0.881
77	0.924
83	0.916
89	0.927
95	0.922
102	0.94
109	0.942
116	0.94

גיל	קוטר
121	0.954
127	0.991
135	0.991
144	0.94
150	0.98
156	1.011
162	1.006
168	1.021

גיל	קוטר
174	1.043
180	1.041
186	1.072
192	1.064
198	1.064
204	1.068

קו הריבועים הפחותים – קוטר לב

לצורך חישוב אומדי הריבועים הפחותים נחשב תחילה את הסכומים הבאים:

$$\sum_{i=1}^{n} x_i = 3039 \qquad \sum_{i=1}^{n} x_i^2 = 456537$$

$$\sum_{i=1}^{n} y_i = 21.638 \qquad \sum_{i=1}^{n} y_i^2 = 21.3538$$

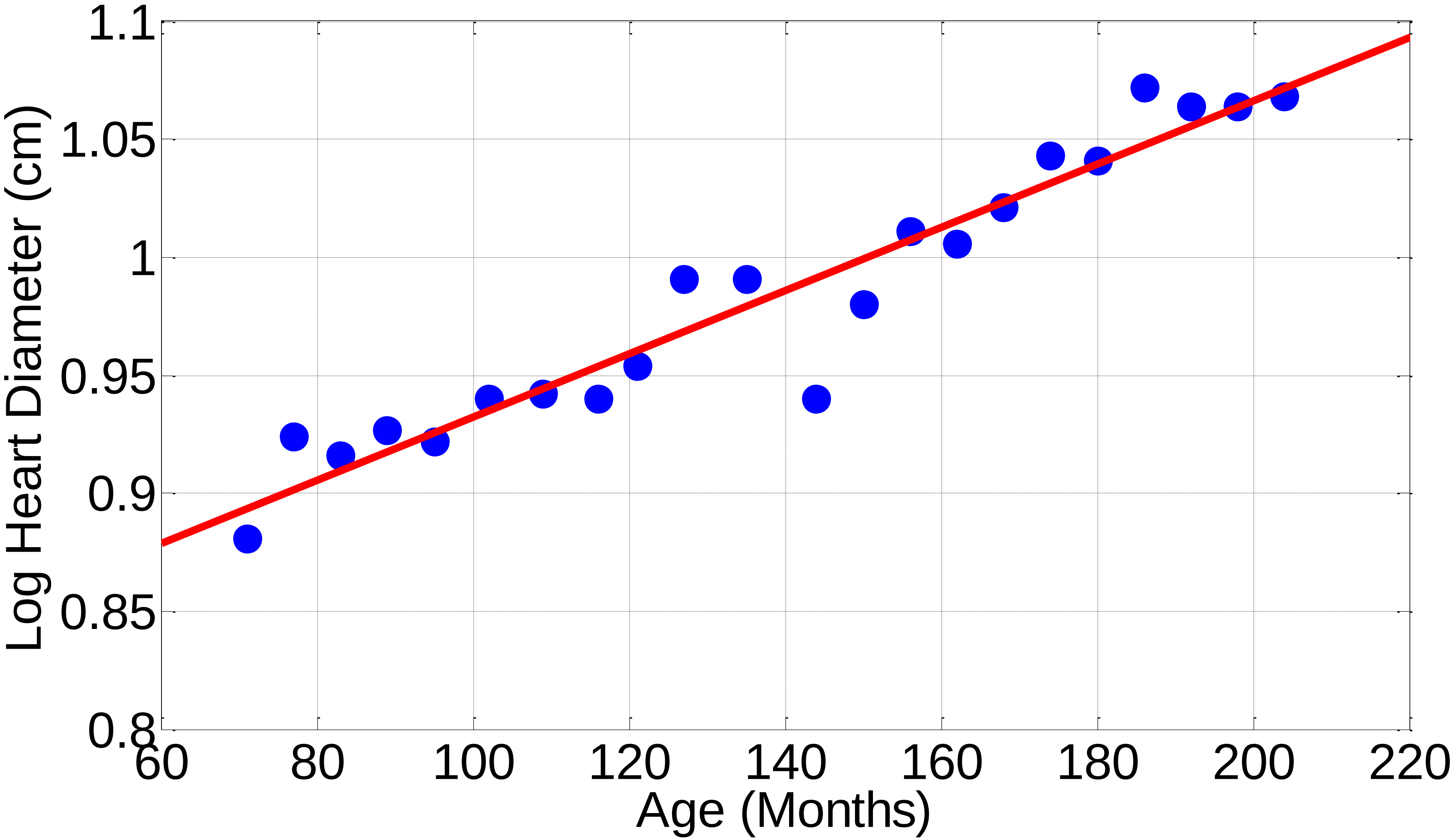
$$\sum_{i=1}^{n} x_i y_i = 3038.156$$

כעת נחשב את אומדי הריבועים הפחותים:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\overbrace{3038.156 - 22\overbrace{(3039/22)}^{\bar{x}=138.14}\overbrace{(21.638/22)}^{\bar{y}=0.98355}}^{SXY=49.161.36}}{\underbrace{456537 - 22(3039/22)^2}_{(n-1)S_{xx}=SXX=36740.59}} = 0.00134$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 0.98355 - 688.347 * 138.14 = 0.799$$

קו הריבועים הפחותים – קוטר לב



הגדרת מטריצות ברגרסיה פשוטה

נגדיר את המטריצות והוקטורים:

$$\underbrace{\mathbf{Y}}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \underbrace{\mathbf{X}}_{n\times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \underbrace{\boldsymbol{\beta}}_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \underbrace{\boldsymbol{\varepsilon}}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

רגרסיה פשוטה בכתיב מטריצות

בעזרת סימנים אלה נכתוב את מודל הרגרסיה בכתיב מטריצות

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

משום ש-

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$= \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{bmatrix}$$

דוגמה לרגרסיה פשוטה במטריצות

הטבלה הבאה כוללת נתונים שנלקחו ממחקר העוסק בהשפעת ריכוז האוזון (x) באוויר על יבול הסויה (y).

X אוזון	0.02	0.07	0.11	0.15
Y יבול	242	237	231	201

דוגמא לכתיבה במטריצות

$$\underbrace{\mathbf{X}}_{4\times 2} = \begin{bmatrix} 1 & 0.02 \\ 1 & 0.07 \\ 1 & 0.11 \\ 1 & 0.15 \end{bmatrix} \quad \underbrace{\mathbf{Y}}_{4\times 1} = \begin{bmatrix} 242 \\ 237 \\ 231 \\ 201 \end{bmatrix}$$

מודל הרגרסיה המרובה

מודל הרגרסיה המרובה מוגדר כ-

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_{i,} \quad \forall \; i = 1, \ldots, n$$

כאשר $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ הינם מקדמי הרגרסיה ו- ε_i הן שגיאות. ההנחות על השגיאות זהות להנחות שהנחנו במודל הליניארי הפשוט.

רגרסיה ליניארית מרובה -דוגמא

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k + \varepsilon$$

משתנה מוסבר \ תלוי

Y
משך זמן ההמתנה בתור בסופר

משתנים מסבירים :

X1 מספר פריטים בעגלה	X2 גיל ומגדר	X3 האם הלקוח מדבר הטלפון	X4 האם יש מוצרים במבצע 1+1

X5 האם יש מוצר שדומה למוצר במבצע	X6 האם הלקוח צריך משלוח	X7 אמצעי התשלום	X8 האם יש חבר מועדון

דוגמה

הטבלה שבשקף הבא מציגה נתונים עבור השנים 1927-1941 ו 1948-1962 לגבי:

- אינדקס מחירי המזון (x_1)
- ההכנסה לנפש (x_2)
- צריכת המזון לנפש (y)

צריכת מזון - נתונים

שנה	מחיר	הכנסה	צריכה	שנה	מחיר	הכנסה	צריכה
27	91.7	57.7	88.9	**48**	105.3	82.1	96.7
28	92	59.3	88.9	**49**	102	83.1	96.7
29	93.1	62	89.1	**50**	102.4	88.6	98
30	90.9	56.3	88.7	**51**	105.4	88.3	96.1
31	82.3	52.7	88	**52**	105	89.1	98.1
32	76.3	44.4	85.9	**53**	102.6	92.1	99.1
33	78.3	43.8	86	**55**	100.8	96.5	99.8
34	84.3	47.8	87.1	**56**	100	99.8	101.5
35	88.1	52.1	85.4	**57**	99.8	99.9	99.9
36	88	58	88.5	**58**	101.2	98.4	99.1
38	83.5	55.9	88.6	**59**	98.8	101.8	101
39	82.4	60.3	91.7	**60**	98.4	101.8	100.7
40	83	64.1	93.3	**61**	98.8	103.1	100.8
41	86.2	73.7	95.1	**62**	98.4	105.5	101

קריטריון הריבועים הפחותים –רגרסיה מרובה

הרעיון הוא להביא למינימום את ריבועי השגיאה:

$$\min Z = \min \sum (y_i - \hat{y})^2 = \sum [y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik})]^2$$

בדומה לשיטה שהוצגה במודל הרגרסיה הפשוטה, נגזור את הפונקציה k+1 פעמים ונשווה ל-0 על מנת לקבל את מערכת המשוואות הנורמליות:

$$\sum y_i = n \cdot b_0 + b_1 \sum x_{i1} + +b_2 \sum x_{i2} + \cdots + +b_p \sum x_{ik}$$

$$\sum x_{i1} \cdot y_i = b_0 \sum x_{i1} + b_1 \sum {x_{i1}}^2 + +b_2 \sum x_{i1} \cdot x_{i2} + \cdots + +b_p \sum x_{i1} \cdot x_{ik}$$

$$\vdots$$

$$\sum x_{ip} \cdot y_i = b_0 \sum x_{ip} + b_1 \sum x_{ip} \cdot x_{i1} + +b_2 \sum x_{ip} \cdot x_{i2} + \cdots + +b_p \sum {x_{ik}}^2$$

כתיבה מטריציונית של המשוואות הנורמליות

ניתן לכתוב את המשוואות הנורמליות שהוצגו בשקף הקודם באופן הבא:

$$\mathbf{X'Y} - \mathbf{X'X\beta} = \mathbf{0}$$

ואת אומדי הריבועים הפחותים:

$$\underbrace{\hat{\boldsymbol{\beta}}}_{(k+1)\times 1} = \underbrace{\underbrace{(\underbrace{\mathbf{X'}}_{(k+1)\times n}\underbrace{\mathbf{X}}_{n\times(k+1)})^{-1}}_{(k+1)\times(k+1)}\underbrace{\mathbf{X'}}_{(k+1)\times n}\underbrace{\mathbf{Y}}_{n\times 1}}_{(k+1)\times 1}$$

כתיב מטריצות לרגרסיה מרובה

מודל הרגרסיה המרובה בכתיב מטריצות מוצג באופן הבא:

$$\underbrace{\mathbf{Y}}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times(k+1)} \underbrace{\boldsymbol{\beta}}_{(k+1)\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{n\times 1}$$

כאשר המטריצות מוגדרות כלהלן:

$$\underbrace{\mathbf{Y}}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \underbrace{\mathbf{X}}_{n\times(k+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad \underbrace{\boldsymbol{\beta}}_{(p+1)\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \underbrace{\boldsymbol{\varepsilon}}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

חישוב האומדים בדוגמה של האוזון והסויה

$$\underbrace{\mathbf{X}}_{4\times 2} = \begin{bmatrix} 1 & 0.02 \\ 1 & 0.07 \\ 1 & 0.11 \\ 1 & 0.15 \end{bmatrix} \qquad \underbrace{\mathbf{Y}}_{4\times 1} = \begin{bmatrix} 242 \\ 237 \\ 231 \\ 201 \end{bmatrix}$$

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0.02 & 0.07 & 0.11 & 0.15 \end{bmatrix} \begin{bmatrix} 1 & 0.02 \\ 1 & 0.07 \\ 1 & 0.11 \\ 1 & 0.15 \end{bmatrix} = \begin{bmatrix} 4 & 0.35 \\ 0.35 & 0.0399 \end{bmatrix}$$

$$\left(\mathbf{X'X}\right)^{-1} = \frac{1}{\det(\mathbf{X'X})} \begin{bmatrix} 0.0399 & -0.35 \\ -0.35 & 4 \end{bmatrix} = \frac{1}{\underbrace{(4*0.0399-.35^2)}_{0.0371}} \begin{bmatrix} 0.0399 & -0.35 \\ -0.35 & 4 \end{bmatrix}$$

אומדים לאוזון וסויה - המשך

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0.02 & 0.07 & 0.11 & 0.15 \end{bmatrix} \begin{bmatrix} 242 \\ 257 \\ 231 \\ 201 \end{bmatrix} = \begin{bmatrix} 911 \\ 76.99 \end{bmatrix}$$

$$\underbrace{\hat{\boldsymbol{\beta}}}_{2\times 1} = (\mathbf{X'X})^{-1}\mathbf{X'Y} = \frac{1}{0.0371} \begin{bmatrix} 0.0399 & -0.35 \\ -0.35 & 4 \end{bmatrix} \begin{bmatrix} 911 \\ 76.99 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 253.434 \\ -293.531 \end{bmatrix}$$

דוגמה –
חיזוי המשכורת החודשית על סמך כמות פרויקטים והותק

במחקר העוסק בזיהוי המשתנים המשפיעים על רמת המשכורת החודשית של מנהלי פרויקטים נבדקו המשתנים ותק (בשנים) וכמות הפרויקטים לעובד לחודש. בידי החוקרים 9 תצפיות:

משכורת חודשית ברוטו	כמות פרויקטים	ותק (שנים)
15000	5	6
6500	1	1
17500	6	15
7000	2	3
8000	5	4
30000	10	20
6000	1	0
8000	5	8
7500	3	4

פתרון

$$Y = \begin{pmatrix} 15000 \\ 6500 \\ 17500 \\ 7000 \\ 8000 \\ 30000 \\ 6000 \\ 8000 \\ 7500 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 6 & 5 \\ 1 & 1 & 1 \\ 1 & 15 & 6 \\ 1 & 3 & 2 \\ 1 & 4 & 5 \\ 1 & 20 & 10 \\ 1 & 0 & 1 \\ 1 & 8 & 5 \\ 1 & 4 & 3 \end{pmatrix}$$

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 1 & 15 & 3 & 4 & 20 & 0 & 8 & 4 \\ 5 & 1 & 6 & 2 & 5 & 10 & 1 & 5 & 3 \end{pmatrix}$$

$$X^T \cdot \mathrm{Y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 1 & 15 & 3 & 4 & 20 & 0 & 8 & 4 \\ 5 & 1 & 6 & 2 & 5 & 10 & 1 & 5 & 3 \end{pmatrix} \cdot \begin{pmatrix} 15000 \\ 6500 \\ 17500 \\ 7000 \\ 8000 \\ 30000 \\ 6000 \\ 8000 \\ 7500 \end{pmatrix} = \begin{pmatrix} 105500 \\ 1106000 \\ 609000 \end{pmatrix}$$

$$X^T \cdot \mathrm{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 1 & 15 & 3 & 4 & 20 & 0 & 8 & 4 \\ 5 & 1 & 6 & 2 & 5 & 10 & 1 & 5 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 6 & 5 \\ 1 & 1 & 1 \\ 1 & 15 & 6 \\ 1 & 3 & 2 \\ 1 & 4 & 5 \\ 1 & 20 & 10 \\ 1 & 0 & 1 \\ 1 & 8 & 5 \\ 1 & 4 & 3 \end{pmatrix} = \begin{pmatrix} 9 & 61 & 38 \\ 61 & 767 & 399 \\ 38 & 399 & 226 \end{pmatrix}$$

M(1,1)=767*226-399*399=14141

M(1,2)=61*226-399*38=-1376

$$\begin{pmatrix} + & - & + \\ - & + & - \\ + & - & + \end{pmatrix} \qquad A^T = \begin{pmatrix} 14141 & 1376 & -4807 \\ 1376 & 590 & -1273 \\ -4807 & -1273 & 3182 \end{pmatrix}$$

$$\det(X^T\mathrm{X}) = 9 \cdot (767 \cdot 226 - 399 \cdot 399) - 61 \cdot (61 \cdot 226 - 399 \cdot 38) + 38 \cdot (61 \cdot 399 - 38 \cdot 767) = 28{,}539$$

$$(X^T\mathrm{X})^{-1} = \frac{1}{\det(X^T\mathrm{X})} \cdot A^T = \frac{1}{28539} \cdot \begin{pmatrix} 14141 & 1376 & -4807 \\ 1376 & 590 & -1273 \\ -4807 & -1273 & 3182 \end{pmatrix} = \begin{bmatrix} 0.495 & 0.048 & -0.168 \\ 0.048 & 0.0206 & -0.044 \\ -0.168 & -0.044 & 0.111 \end{bmatrix}$$

$$\hat{\beta} = (X^T\mathrm{X})^{-1} \cdot (X^T\mathrm{Y}) = \begin{pmatrix} 0.495 & 0.048 & -0.168 \\ 0.048 & 0.0206 & -0.044 \\ -0.168 & -0.044 & 0.111 \end{pmatrix} \cdot \begin{pmatrix} 105500 \\ 1106000 \\ 609000 \end{pmatrix} = \begin{bmatrix} 3022.83 \\ 786.62 \\ 798.23 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

על כן, משוואת הניבוי:

$$\hat{Y} = 3022.83 + 786.62X_1 + 798.23X_2$$