

Biostatistics

Week #17

6/14/2022



Chapter 17 – Correlation & Regression

- Correlation (Pearson's correlation coefficient)
- Linear Regression
- Multiple Regression

Introduction

- To determine whether there is an association between two variables (one independent and one dependent)
- If so, ***what is the association?***
- Can we use it to ***predict*** the weight of a male bear given his body length?

Lengths and Weights of Male Bears

x Length	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight	80	344	416	348	262	360	332	34

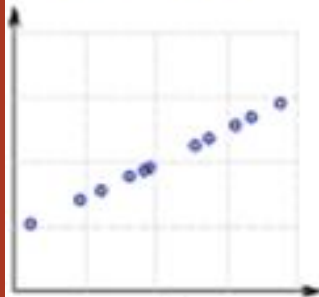
Correlation

- A “correlation” can help determining whether there is a “statistically significant” association between two variables.
- A scatter plot can help visually assessing whether the paired data (x, y) might be correlated.
- Such correlation could be “linear” or in other nonlinear forms (such as exponential, etc.)

Linear correlation

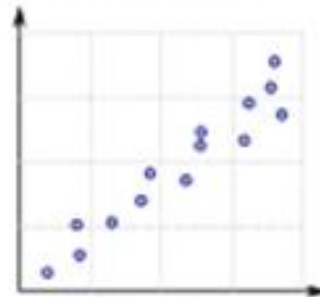
- The *linear correlation coefficient r* measures the strength of the linear association between the paired x- and y-quantitative values in a sample.
- It is sometimes called *Pearson product moment correlation coefficient*.
- r ranges between -1 and 1.

Perfect
Positive
Correlation



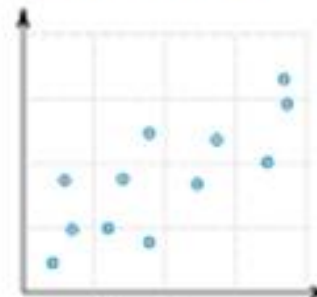
1

High
Positive
Correlation



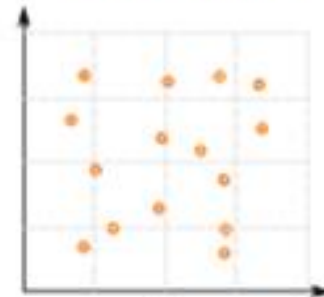
0.8

Low
Positive
Correlation



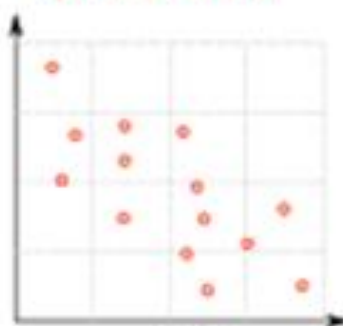
0.3

No
Correlation



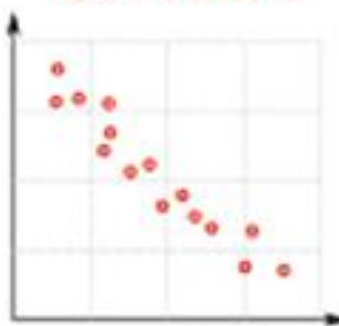
0

Low
Negative
Correlation



-0.3

High
Negative
Correlation



-0.8

Perfect
Negative
Correlation



-1

Basic requirements before computing r

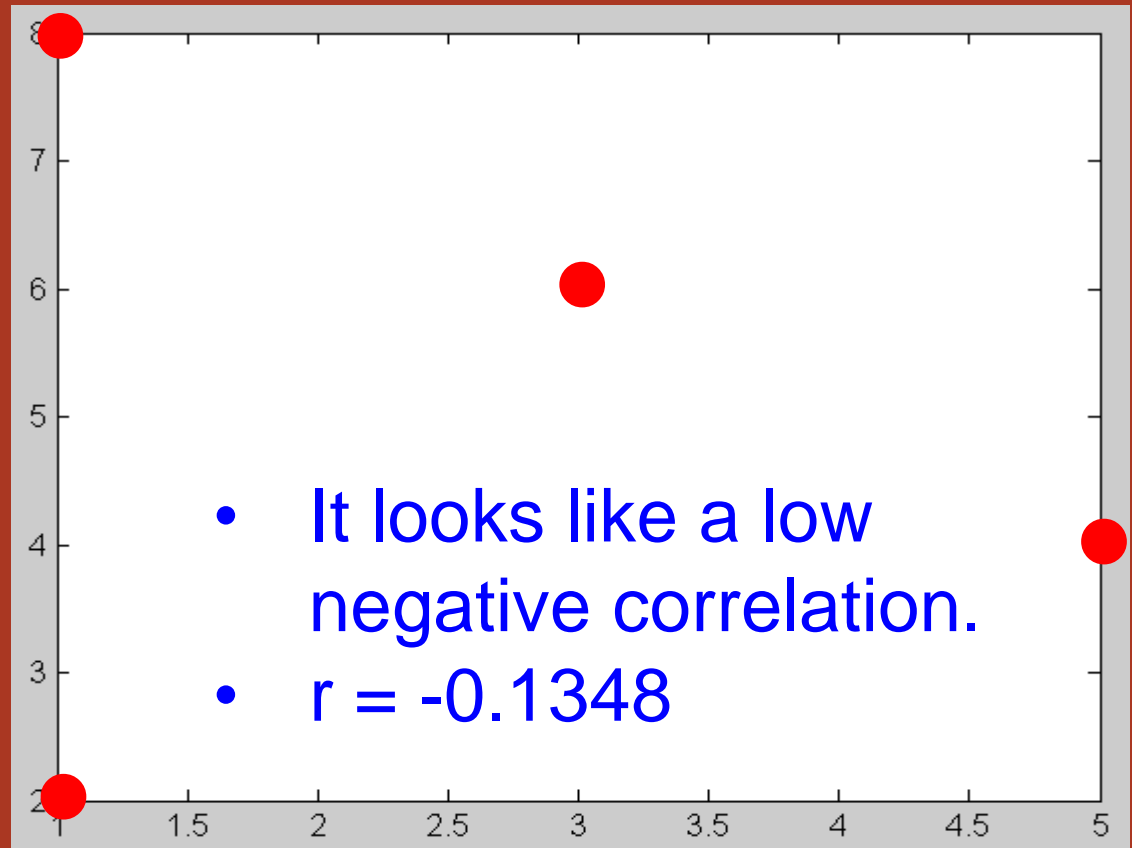
1. paired data (x, y) are randomly sampled.
2. visual scatter plot be approximately a straight line.
3. outliers be firstly removed

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2] [n(\sum y^2) - (\sum y)^2]}}$$

Example 1

Given the following 4 paired data, compute the correlation coefficient r .

x	1	1	3	5
y	2	8	6	4




```
>> x=[1 1 3 5];y=[2 8 6 4];n=4;  
>> r=(n*sum(x.*y)-  
sum(x)*sum(y))/sqrt((n*sum(x.*x)-  
sum(x)^2)*(n*sum(y.*y)-sum(y)^2))
```

r = -0.1348

```
>>
```

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2] [n(\sum y^2) - (\sum y)^2]}}$$

```
>> x=[1 1 3 5];y=[2 8 6 4];n=4;
```

```
>> r=(n*sum(x.*y)-  
sum(x)*sum(y))/sqrt((n*sum(x.*x)-  
sum(x)^2)*(n*sum(y.*y)-sum(y)^2))
```

```
r = -0.1348
```

```
>>
```

Pay special attention to the usage of computing those summations.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2] [n(\sum y^2) - (\sum y)^2]}}$$

```
>> x=[1 1 3 5];y=[2 8 6 4];n=4;  
>> r=(n*sum(x.*y)-  
sum(x)*sum(y))/sqrt((n*sum(x.*x)-  
sum(x)^2)*(n*sum(y.*y)-sum(y)^2))
```

```
r = -0.1348
```

```
>>
```

Pay special attention to the usage of computing those summations.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Interpreting r

- r is between -1 (perfect negative correlation) and +1 (perfect positive correlation).
- $r = 0$ means no correlation.
- Then what defines a “strong” correlation?
- The absolute value of r should be no less than ***a critical value***?
- Is r a random variable having its own probability density function?

Hypothesis testing for r

- H_0 : no significant linear correlation
- Test statistic t :
- This is a 2-tailed t test
- $DF = n-2$
- $\alpha = 0.05$ (usually)
- p-value can be computed based on computed t on a t distribution of specific DF.
- Reject if $p \leq 0.05$.

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Back to example 1 (n=4)

- A critical t value that cuts 0.025 off the left tail of $t_{DF=2}$ is “ $\text{tinv}(0.025, 2) = -4.3027$ ”.
- The t-statistic computed based on previously computed $r = -0.1348$ now becomes -0.1925.
- -0.1925 is not as extreme or more extreme than -4.3027. We thus do not reject the null hypothesis of no significant linear correlation. 【There exists NO significant correlation~~~】
- P-value = $2 * \text{tcdf}(-0.1925, 2) = 0.8652$, much greater than $\alpha = 0.05$.

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

```
>> x=[1 1 3 5];y=[2 8 6 4];  
>> n=4;  
>> r=(n*sum(x.*y)-  
sum(x)*sum(y))/sqrt((n*sum(x.*x)-  
sum(x)^2)*(n*sum(y.*y)-sum(y)^2))
```

```
r = -0.1348
```

```
>> t=r/sqrt((1-r^2)/(n-2))
```

```
t = -0.1925
```

```
>> 2*tcdf(t,n-2)
```

```
ans = 0.8652
```

***Conclusion – No
correlation between the two
variables.***

Example 2

- Find the linear correlation coefficient r for the following data.
- Determine whether the correlation is significant or not by computing a critical r value and a p -value.

Lengths and Weights of Male Bears

x Length	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight	80	344	416	348	262	360	332	34

- The computed $r = \underline{0.8974}$.
- The computed t-statistic = 4.9807.
- $DF=8-2=6$
- A critical r cutting off 0.025 of the right tail of $t_{DF=6}$ is “ $\text{tinv}(0.975, 6)=2.4469$ ”. This is smaller than 4.9807. So our t-statistic is more extreme than expected.
- P-value = “ $2*(1-\text{tcdf}(4.9807,6))=\underline{0.0025}$ ”.
- We thus reject the null hypothesis, suggesting a significant linear correlation exists between the length and weight for male bears.

```
>> x=[53 67.5 72 72 73.5 68.5 73 37]; y=[80 344  
416 348 262 360 332 34];n=8;
```

```
>> r=(n*sum(x.*y)-  
sum(x)*sum(y))/sqrt((n*sum(x.*x)-  
sum(x)^2)*(n*sum(y.*y)-sum(y)^2))
```

```
r = 0.8974
```

```
>> t=r/sqrt((1-r^2)/(n-2))
```

```
t = 4.9807
```

```
>> 2*(1-tcdf(t,6))
```

```
ans = 0.0025
```

```
>>
```

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Using MATLAB's “corrcoef” function

```
>> x=[53 67.5 72 72 73.5 68.5 73 37];  
>> y=[80 344 416 348 262 360 332 34];  
>> [R, P] = corrcoef(x,y)
```

R =

1.0000	<u>0.8974</u>
0.8974	1.0000

← Pearson coefficient

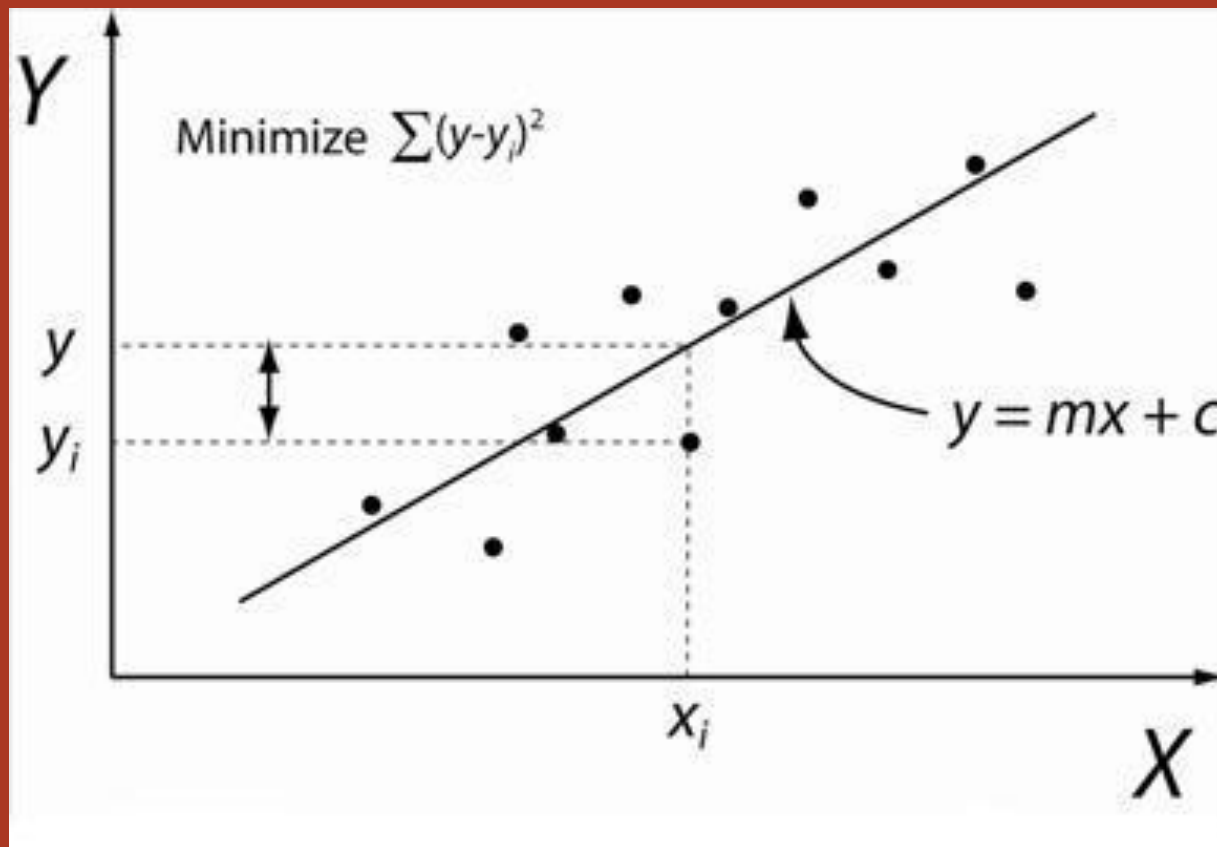
P =

1.0000	<u>0.0025</u>
0.0025	1.0000

← P-value in supporting
the linear correlation
at $\alpha=0.05$.

Linear Regression

- To find a graph and an equation of the straight line that represents the association.
- The straight line is called “regression line”.
- The equation is called “regression equation”.



It's all about finding the slope m and the y-intercept c of the straight line.

Example 3

- Find the regression equation for the following data.
- Predict the weight of a bear with $x = 71.0$.

Lengths and Weights of Male Bears

x Length	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight	80	344	416	348	262	360	332	34

MATLAB's "polyfit" (based on minimizing the least-squares of the errors) will serve the purpose.

```
>> x=[53 67.5 72 72 73.5 68.5 73 37];  
>> y=[80 344 416 348 262 360 332 34];  
>> polyfit(x,y,1)  
ans =  
    9.6598 -351.6599  
>>
```

The equation is $y = 9.6598x - 351.6599$

MATLAB's "polyval" can be used to evaluate a value of a polynomial function.

```
>> x=[53 67.5 72 72 73.5 68.5 73 37];  
>> y=[80 344 416 348 262 360 332 34];  
>> polyfit(x,y,1)  
ans =  
    9.6598 -351.6599  
>> polyval(polyfit(x,y,1), 71.0)  
ans = 334.1849
```

```
>> A male bear of length 71.0 in  
would weigh 334.1849 pounds.
```


Multiple regression

- Two or more independent variables.

Data from Male Bears

y Weight	80	344	416	348	262	360	332	34
x2 Age	19	55	81	115	56	51	68	8
x3 Head L	11.0	16.5	15.5	17.0	15.0	13.5	16.0	9.0
x4 Head W	5.5	9.0	8.0	10.0	7.5	8.0	9.0	4.5
x5 Neck	16.0	28.0	31.0	31.5	26.6	27.0	29.0	13.0
x6 Length	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
x7 Chest	26	45	54	49	41	49	44	19

$$y = b1 + b2*x2 + b3*x3 + \dots + b6*x6 + b7*x7$$

Example 4

- Find b_1 , b_3 and b_6 .

Data from Male Bears

y Weight	80	344	416	348	262	360	332	34
x2 Age	19	55	81	115	56	51	68	8
x3 Head L	11.0	16.5	15.5	17.0	15.0	13.5	16.0	9.0
x4 Head W	5.5	9.0	8.0	10.0	7.5	8.0	9.0	4.5
x5 Neck	16.0	28.0	31.0	31.5	26.6	27.0	29.0	13.0
x6 Length	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
x7 Chest	26	45	54	49	41	49	44	19

$$y = b_1 + b_3 * x_3 + b_6 * x_6$$

$$b1 + b3*x3 + b6*x6 = y$$

$$b1 + 11.0*b3 + 53.0*b6 = 80$$

$$b1 + 16.5*b3 + 67.5*b6 = 344$$

$$b1 + 15.5*b3 + 72.0*b6 = 416$$

$$b1 + 17.0*b3 + 72.0*b6 = 348$$

$$b1 + 15.0*b3 + 73.5*b6 = 262$$

$$b1 + 13.5*b3 + 68.5*b6 = 360$$

$$b1 + 16.0*b3 + 73.0*b6 = 332$$

$$b1 + 9.0*b3 + 37.0*b6 = 34$$

$$\begin{bmatrix} 1 & 11.0 & 53.0 \\ 1 & 16.5 & 67.5 \\ 1 & 15.5 & 72.0 \\ 1 & 17.0 & 72.0 \\ 1 & 15.0 & 73.5 \\ 1 & 13.5 & 68.5 \\ 1 & 16.0 & 73.0 \\ 1 & 9.0 & 37.0 \end{bmatrix} \begin{bmatrix} b1 \\ b3 \\ b6 \end{bmatrix} = \begin{bmatrix} 80 \\ 344 \\ 416 \\ 348 \\ 262 \\ 360 \\ 332 \\ 34 \end{bmatrix}$$

*This is called an **over-determined system**. We have 8 equations, more than needed to solve 3 unknowns ($b1$, $b3$ and $b6$).*

or **$AX = y$**


$$AX = y$$

$$[8 \text{ by } 3][3 \text{ by } 1] = [8 \text{ by } 1]$$

- The problem is – matrix A is not square. We cannot find its inverse and solve the equation as $X=A^{-1}y$.

Pseudo-inverse of matrix A

$$X = \text{pinv}(A)y$$

$$[3 \text{ by } 1] = [3 \text{ by } \underline{8}] [8 \text{ by } 1]$$

- I need to have a “3 by 8” matrix which serves like an inverse of A. We call it a pseudo-inverse of matrix A, or $\text{pinv}(A) = (A^t A)^{-1} A^t$.
- See Appendix for the definition of $\text{pinv}(A)$

```
>> y=[80 344 416 348 262 360 332 341];  
>> x3=[11 16.5 15.5 17 15 13.5 16 9];  
>> x6=[53 67.5 72 72 73.5 68.5 73 37];  
>> A=[ones(size(x3)) x3 x6]
```

Note that y,
x3 and x6
must be
column
vectors.

A =

1.0000	11.0000	53.0000
1.0000	16.5000	67.5000
1.0000	15.5000	72.0000
1.0000	17.0000	72.0000
1.0000	15.0000	73.5000
1.0000	13.5000	68.5000
1.0000	16.0000	73.0000
1.0000	9.0000	37.0000

[8 by 3]

>> A'*A

ans =

1.0e+004 *

0.0008 0.0114 0.0517

0.0114 0.1667 0.7565

0.0517 0.7565 3.4526

$$\begin{matrix} A' & * & A \\ [3 \text{ by } 8] & [8 \text{ by } 3] & = [3 \text{ by } 3] \end{matrix}$$

$$\text{pinv}(A) = (A^t A)^{-1} A^t$$

>> pinvA=inv(ans)*A'

$$\begin{matrix} (A^t A)^{-1} & * & A^t \\ [3 \text{ by } \underline{8}] & = & [3 \text{ by } 3] [3 \text{ by } 8] \end{matrix}$$

pinvA =

0.8763 -0.2719 -0.2571 -0.4628 -0.2293 0.1123 -0.3528 1.5853

-0.0983 0.1962 -0.0209 0.1501 -0.1123 -0.1686 0.0132 0.0405

0.0100 -0.0370 0.0105 -0.0239 0.0302 0.0373 0.0045 -0.0315

>> b=pinvA*y

b =

-374.3035

18.8204

5.8748

$$[3 \text{ by } 1] = [3 \text{ by } \underline{8}] [8 \text{ by } 1]$$

$$y = -374.3035 + 18.8204*x3 + 5.8748*x6$$

>>

MATLAB's "regress"

```
>> y=[80 344 416 348 262 360 332 34]',
```

```
>> x3=[11 16.5 15.5 17 15 13.5 16 9]';
```

```
>> x6=[53 67.5 72 72 73.5 68.5 73 37]';
```

```
>> A=[ones(size(x3)) x3 x6];
```

```
>> b=regress(y, A)
```

```
b =
```

```
-374.3035
```

```
18.8204
```

```
5.8748
```

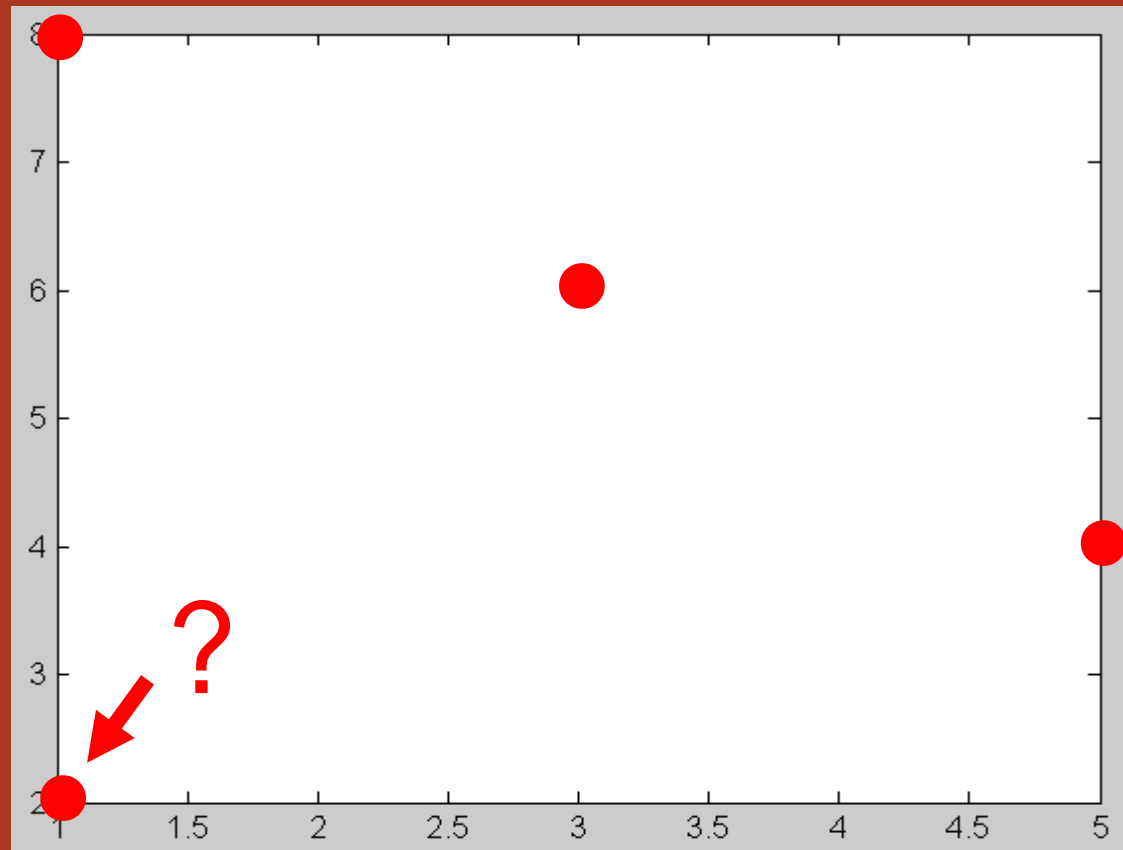
```
>>
```

Note that y, x3 and x6 must be column vectors in order to build the 8x3 matrix A.

In class practice – 1

- What if we treat the first data point in Example 1 as an outlier?

x	1	1	3	5
y	2	8	6	4



Cont'd

- Compute the correlation coefficient r .
- Compute the corresponding t-statistic t .
- Compute the p-value for the null hypothesis that the two variables are not correlated.
- Your conclusion?


In class practice – 2

- Find b_1 , b_2 and b_6 .

Data from Male Bears

y Weight	80	344	416	348	262	360	332	34
x2 Age	19	55	81	115	56	51	68	8
x3 Head L	11.0	16.5	15.5	17.0	15.0	13.5	16.0	9.0
x4 Head W	5.5	9.0	8.0	10.0	7.5	8.0	9.0	4.5
x5 Neck	16.0	28.0	31.0	31.5	26.6	27.0	29.0	13.0
x6 Length	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
x7 Chest	26	45	54	49	41	49	44	19

$$y = b_1 + b_2 * x_2 + b_6 * x_6$$



APPENDIX – LEAST SQUARE METHOD FROM LINEAR ALGEBRA

Least-Squares Curves

- A system $A\mathbf{x} = \mathbf{y}$ of n equations in n variables, where A is invertible, has the unique solution $\mathbf{x} = A^{-1}\mathbf{y}$.
- However, if the system has n equations and m variables, with $n > m$, the system does not, in general, have a solution and it is said to be **over-determined**.
- A is not a square matrix, thus A^{-1} does not exist.
- A matrix called the **pseudoinverse** of A , denoted $\text{pinv}(A)$, leads to a least-squares solution $\mathbf{x} = \text{pinv}(A)\mathbf{y}$ for an over-determined system.
- This is not a true solution, but in some sense the closest we can get in order to have a true solution.

DEFINITION :

Let A be a matrix, then the matrix $(A^t A)^{-1} A^t$ is called a **pseudoinverse** of A and is denoted $\text{pinv}(A)$.

Example 1 Find the pseudoinverse of $A =$

$$\begin{bmatrix} 1 & 2 \\ -1 & 3 \\ 2 & 4 \end{bmatrix}$$

Solution

$$A^t A = \begin{bmatrix} 1 & -1 & 2 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ -1 & 3 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 6 & 7 \\ 7 & 29 \end{bmatrix}$$

$$(A^t A)^{-1} = \frac{1}{|A^t A|} \text{adj}(A^t A) = \frac{1}{125} \begin{bmatrix} 29 & -7 \\ -7 & 6 \end{bmatrix}$$

$$\text{pinv}(A) = (A^t A)^{-1} A^t = \frac{1}{125} \begin{bmatrix} 29 & -7 \\ -7 & 6 \end{bmatrix} \begin{bmatrix} 1 & -1 & 2 \\ 2 & 3 & 4 \end{bmatrix} = \frac{1}{25} \begin{bmatrix} 3 & -10 & 6 \\ 1 & 5 & 2 \end{bmatrix}$$

System of Equations $A\mathbf{x} = \mathbf{y}$

$$A\mathbf{x} = \mathbf{y} \quad \mathbf{x} = \text{pinv}(A)\mathbf{y}$$

system

least-squares solution

Let $A\mathbf{x} = \mathbf{y}$ be a system of n linear equations in m variables with $n > m$, where A is of rank m .

- (1) This system has a least-squares solution.
- (2) If the system has a unique solution, the least –squares solution is that unique solution.
- (3) If the system is over-determined, the least-squares solution is the closest we can get to a true solution.
- (4) The system cannot have many solutions.

Example 2

Find the least-squares solution to the following over-determined system of equations. Sketch the solution.

Solution

We have

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \\ 2 & 3 \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} 6 \\ 3 \\ 9 \end{bmatrix}$$

$$x + y = 6$$

$$-x + y = 3$$

$$2x + 3y = 9$$

$$(m=3, n=2)$$

$$A^t A = \begin{bmatrix} 1 & -1 & 2 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 6 & 6 \\ 6 & 11 \end{bmatrix}$$

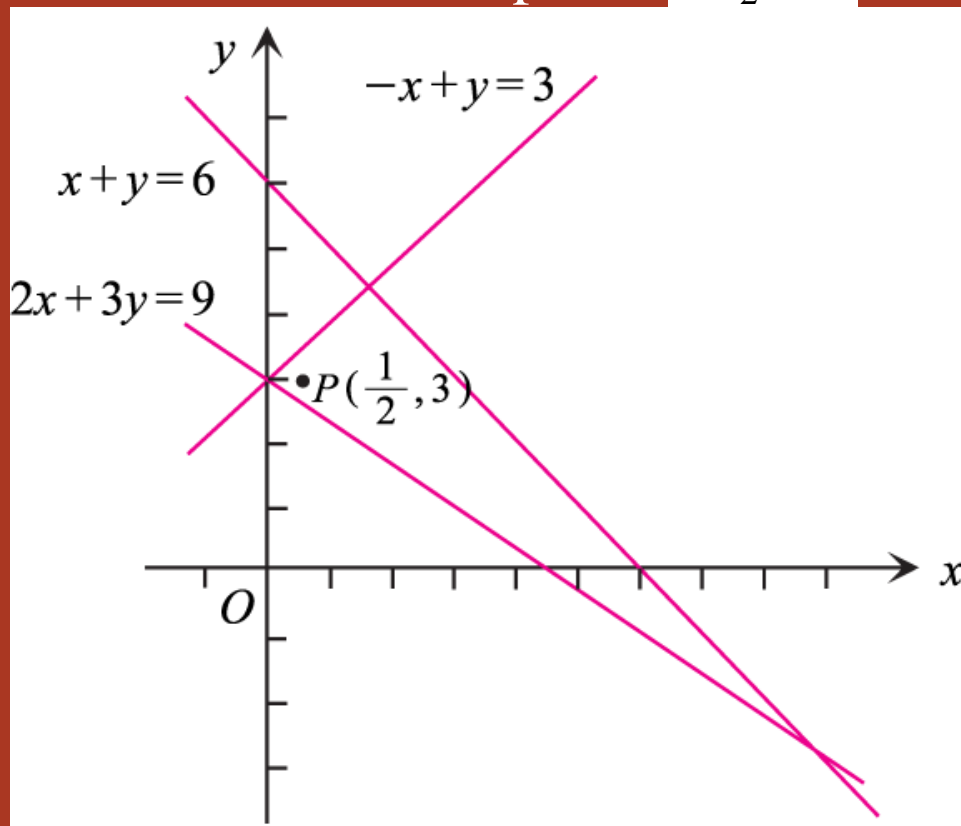
$$(A^t A)^{-1} = \frac{1}{|A^t A|} \text{adj}(A^t A) = \frac{1}{30} \begin{bmatrix} 11 & -6 \\ -6 & 6 \end{bmatrix}$$

$$\text{pinv}(A) = (A^t A)^{-1} A^t = \frac{1}{30} \begin{bmatrix} 11 & -6 \\ -6 & 6 \end{bmatrix} \begin{bmatrix} 1 & -1 & 2 \\ 1 & 1 & 3 \end{bmatrix} = \frac{1}{30} \begin{bmatrix} 5 & -17 & 4 \\ 0 & 12 & 6 \end{bmatrix}$$

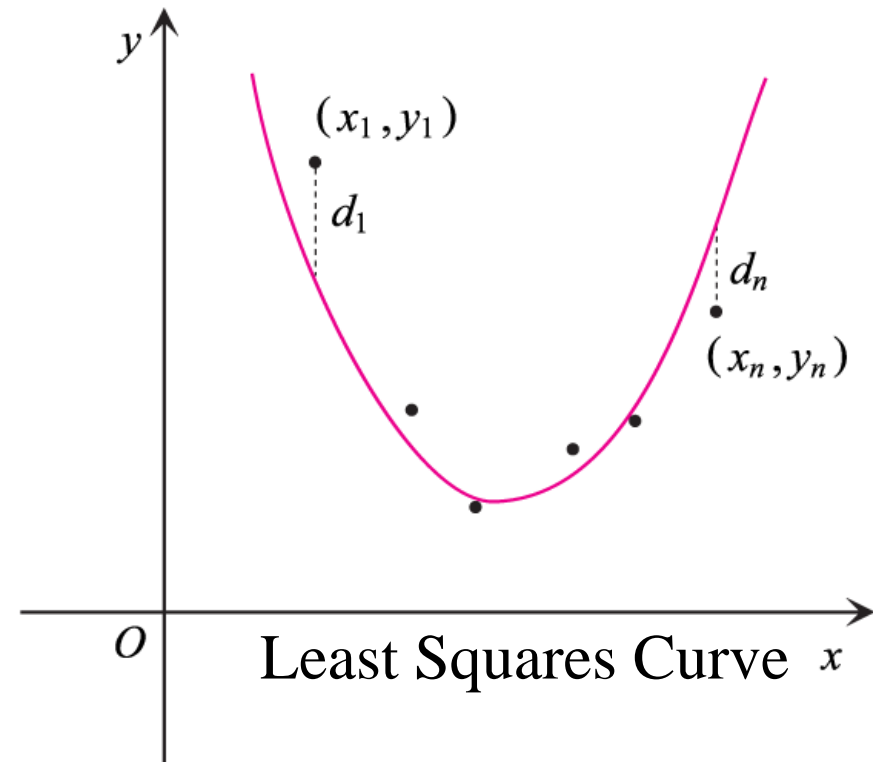
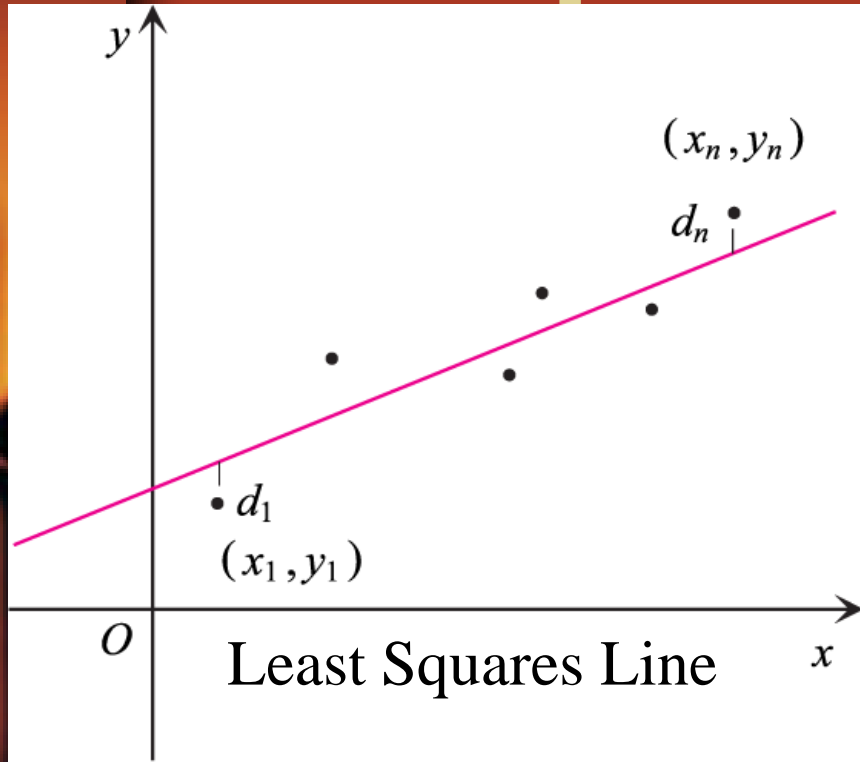
Then the least-squares solution is

$$\text{pinv}(A)\mathbf{y} = \frac{1}{30} \begin{bmatrix} 5 & -17 & 4 \\ 0 & 12 & 6 \end{bmatrix} \begin{bmatrix} 6 \\ 3 \\ 9 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 3 \end{bmatrix}$$

The solution is shown below as point $P(\frac{1}{2}, 3)$.



Least Squares Curves



The least squares line or curve can be found by solving a over-determined system. This is found by minimizing the sum of $d_1^2 + d_2^2 + \dots + d_n^2$

That's where we get the name "least squares" from.

Example 3

Find the least-squares line for the following data points.

$$(1, 1), (2, 2.4), (3, 3.6), (4, 4)$$

Solution

Let the equation of the line be $y = a + bx$. Substituting for these points into the equation, we get an over-determined system:

$$\begin{aligned}a + b &= 1 \\a + 2b &= 2.4 \\a + 3b &= 3.6 \\a + 4b &= 4\end{aligned}$$

To solve for the least-squares solution, we have

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2.4 \\ 3.6 \\ 4 \end{bmatrix}$$

Thus

$$\text{pinv}(A) = (A^t A)^{-1} A^t = \frac{1}{20} \begin{bmatrix} 20 & 10 & 0 & -10 \\ -6 & -2 & 2 & 6 \end{bmatrix}$$

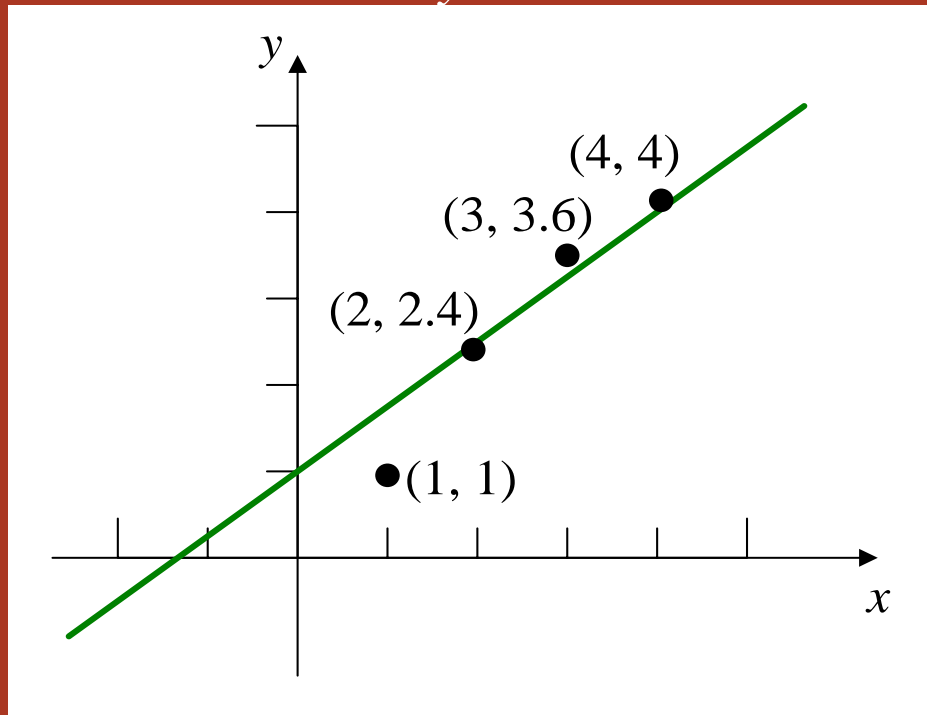
And the least-squares solution is

$$[(A^t A)^{-1} A^t] \mathbf{y} = \frac{1}{20} \begin{bmatrix} 20 & 10 & 0 & -10 \\ -6 & -2 & 2 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 2.4 \\ 3.6 \\ 4 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 1.02 \end{bmatrix}$$

Thus $a = 0.2$, $b = 1.02$.

And the equation is

$$y = 0.2 + 1.02x$$



Example 4

Find the least-squares parabola for the following data points.

$(1, 7), (2, 2), (3, 1), (4, 3)$

Solution

Let the parabola be $y = a + bx + cx^2$. Substituting data points:

$$a + b + c = 7$$

$$a + 2b + 4c = 2$$

$$a + 3b + 9c = 1$$

$$a + 4b + 16c = 3$$

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} 7 \\ 2 \\ 1 \\ 3 \end{bmatrix}$$

We have

$$\text{pinv}(A) = (A^t A)^{-1} A^t = \frac{1}{20} \begin{bmatrix} 45 & -15 & -25 & 15 \\ -31 & 23 & 27 & -19 \\ 5 & -5 & -5 & 5 \end{bmatrix}$$

Finally we have the solution

$$[(A^t A)^{-1} A^t] \mathbf{y} = \frac{1}{20} \begin{bmatrix} 45 & -15 & -25 & 15 \\ -31 & 23 & 27 & -19 \\ 5 & -5 & -5 & 5 \end{bmatrix} \begin{bmatrix} 7 \\ 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 15.25 \\ -10.05 \\ 1.75 \end{bmatrix}$$

Thus $a = 15.25$, $b = -10.05$,
 $c = 1.75$.

Or

$$y = 15.25 - 10.05x + 1.75x^2$$

