Regression Analysis

ECE 341: Random Processes

Lecture #19

note: All lecture notes, homework sets, and solutions are posted on www.BisonAcademy.com

Linear Estimation of Y given X:

Problem: Given measurement Y, estimate X.

- You want to know something that is difficult to measure. You estimate this based upon something that is easier to measure.
  - Fan speed ≈ thrust for a jet engine (GE)
  - Pressure drop ≈ thrust (Pratt & Whitney)

Since the estimate is different from the 'true' value, denote

$\hat{x}$ The estimate of x

x The 'true' value of x

$\bar{x}$ The mean of x

B Basis matrix: functions of x

Form an estimate based upon Y using a linear curve fit:

$$\hat{y} = ax + b$$

Least Squares

Procedure to find the parameters 'a' and 'b' given n data points:

Step 1) Write this in matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

or

$$Y = BA$$

You can't invert matrix B since it's not square. To make it square, multiply by B transpose:

$$B^T Y = B^T B \cdot A$$

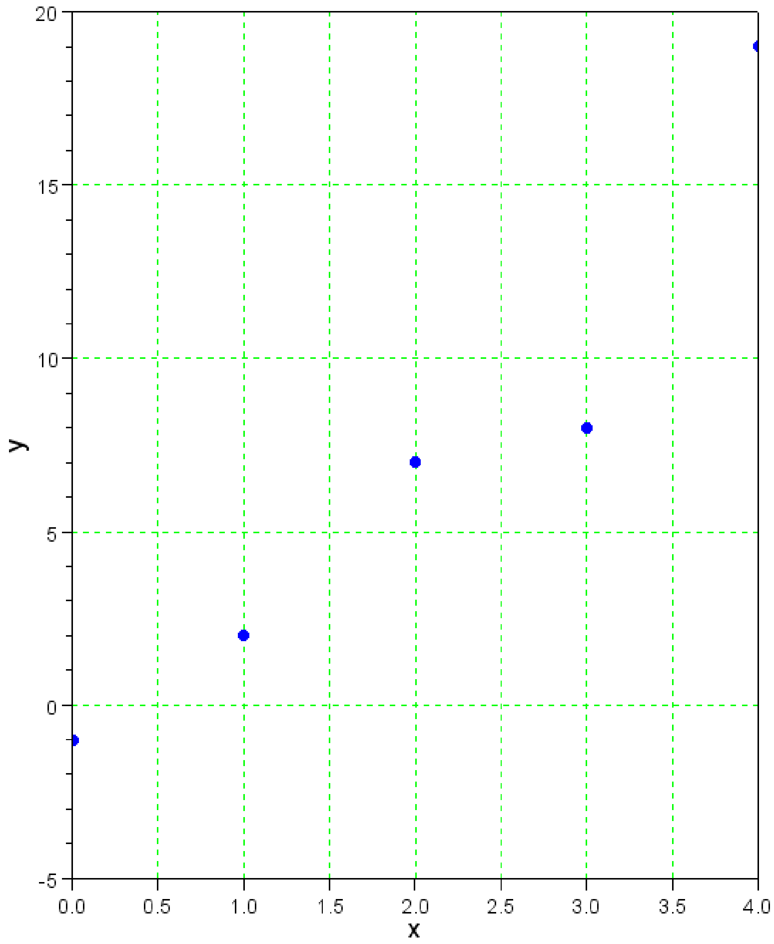
$B^T B$ is square and is usually invertable. Solve for A:

$$A = \left(B^T B\right)^{-1} B^T Y$$

This is the least squares solution for a and b.

Example: Find the least squares curve fit for the following data points (x,y)

x	y
0.	-1.
1.	2.
2.	7.
3.	8.
4.	19.

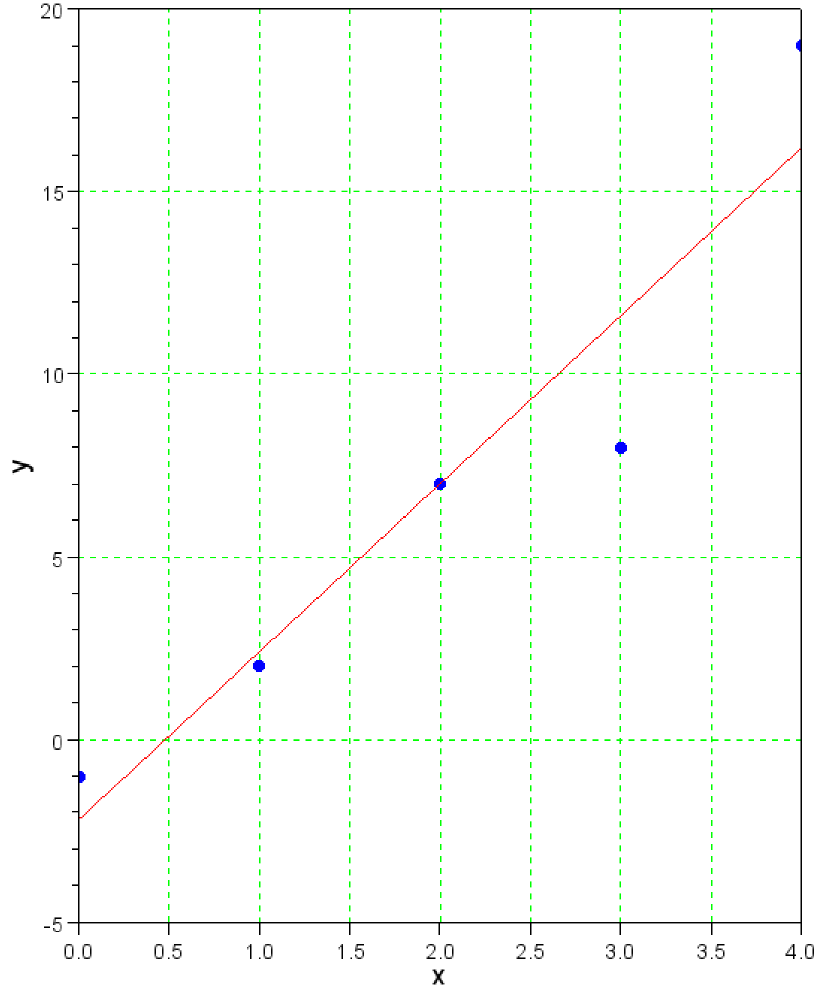


Solution: Create matrix B that defines your basis functions:

```
B = [x, x.^0]
     0.     1.
     1.     1.
     2.     1.
     3.     1.
     4.     1.
```

Determine 'a' and 'b'

```
A = inv(B'*B)*B'*y
     4.6       times x
   - 2.2       times 1

plot(x, y, 'b.', x, y, 'r-');
```



So, the least squares estimate for y(x) is:

$$\hat{y} \approx 4.6x - 2.2$$

This minimizes the sum-squared error

$$J = \sum (y_i - \hat{y}_i)^2$$

Weighted Least squares:

If you 'trust' some data points more than others, you can weight the data. For example, suppose you weight (trust) the 4th data point 10.6 times more than the rest.

```
 x    y    q (weight)
 0.  - 1.    1
 1.    2.    1
 2.    7.    1
 3.    8.    10.6
 4.    19.   1
```

Create a diagonal matrix, Q, which has the weight for each element:

```
Q = diag([1,1,1,10.6,1])

    1.    0.    0.    0.      0.
    0.    1.    0.    0.      0.
    0.    0.    1.    0.      0.
    0.    0.    0.    10.6    0.
    0.    0.    0.    0.      1.
```

Return to the equation for X and Y in matrix form:

$$Y = B\ A$$

Multiply by Q

$$QY = QB\ A$$

Multiply by X transpose

$$B^T\ QY = B^TQB\ A$$

Invert

$$(B^TQB)^{-1}\ B^T\ QY = A$$

The results is the least squares solution with weighting Q:

$$J = \sum q_i (y_i - \hat{y}_i)^2$$

Going back to our example:

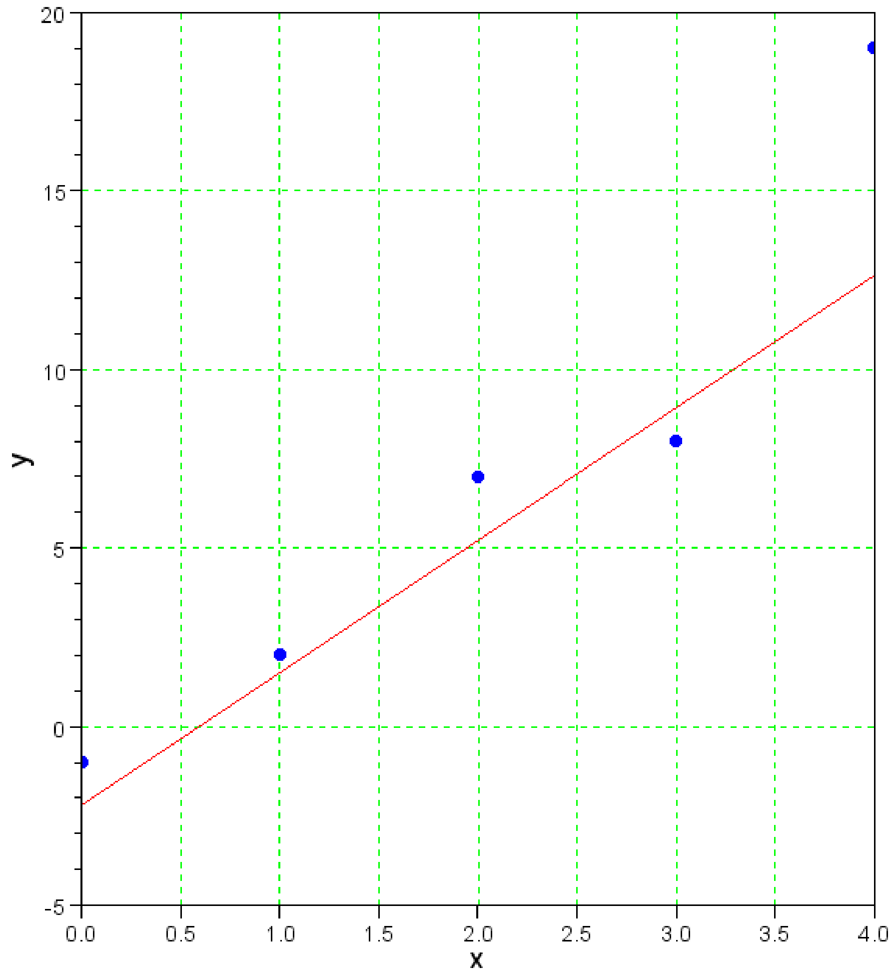
```
-->Q = diag([1,1,1,10.6,1])
-->A = inv(B'*Q*B)*B'*Q*Y

    3.7092784
  - 2.2
```

so now the estimate for y should be:

$$\hat{y} = 3.70927x - 2.2$$

Checking by plotting this vs. your data:

```
-->y1 = 3.7092784*x1 - 2.2;
-->plot(x,y,'.',x1,y1,'-r')

-->xlabel('x')
-->ylabel('y')
```



Note that the line is closer to the 4th data point (3,8) due to its weight of 10.6.

Covariance and Correlation Coefficient

The correlation between X and Y tells you how closely the two are related

- Correlation of zero means they are independent
- Correlation of +1.000 means that as X increases, Y increases.
- Correlation of -1.000 means that as X increases, Y decreases.

Correlation doesn't care about cause and affect: it just tells you whether the two behave the same way.

- Useful in jet engines: measure something highly correlated with thrust
- Useful in Wall Street: measure something that his highly correlated with stock prices 1 year in the future.

To determine the correlation coefficient, you first need to determine the covariance between X and Y.

Covariance:

The covariance between X and Y is defined as

$$Cov[X,Y] = E[(x-\bar{x})(y-\bar{y})]$$

Doing some algebra

$$Cov[X,Y] = E[(x-\bar{x})(y-\bar{y})]$$

$$= E[xy] - \bar{x} \cdot \bar{y}$$

The correlation coefficient is defined as

$$\rho_{X,Y} = \frac{Cov[X,Y]}{\sqrt{\sigma_X^2 \cdot \sigma_Y^2}}$$

- $\rho = \pm 1$ x and y are 100% correlated. If you know x you know y with no error.
- $\rho = 0$ x and y have no correlation. Knowing x tells you nothing about y.

Some other useful relationships are

1st moment (m1)

$$m_1 = mean(x)$$

2nd moment (m2)

$$m_2 = mean(x^2)$$

Variance

$$\sigma^2 = m_2 - m_1^2$$

Covariance:

$$Cov(X, Y) = mean(xy) - mean(x)\ mean(y)$$

Correlation coefficient

$$\rho_{X,Y} = \left(\frac{Cov(X,Y)}{\sqrt{\sigma_x^2 \sigma_y^2}} \right)$$

Examples: Let

- x0 be a variable in the range if (0,10)
- n be noise: random variable with a uniform distribution over (0,10)

Let x be 0% to 100% noise

$$x = \alpha x_0 + (1 - \alpha)\eta$$

Let y be related to x as

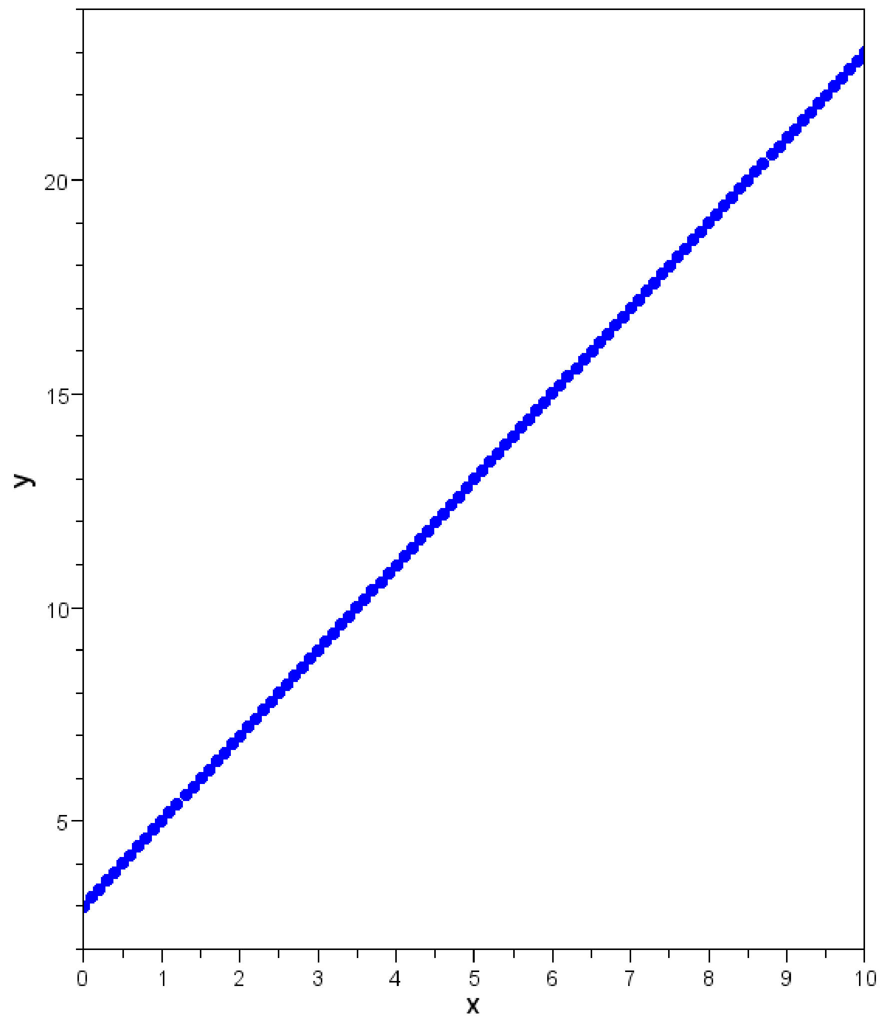
$$y = 2x + 3$$

Determine how the correlation coefficient varies with alpha.

No Noise

- $\rho^2 = 1.000$

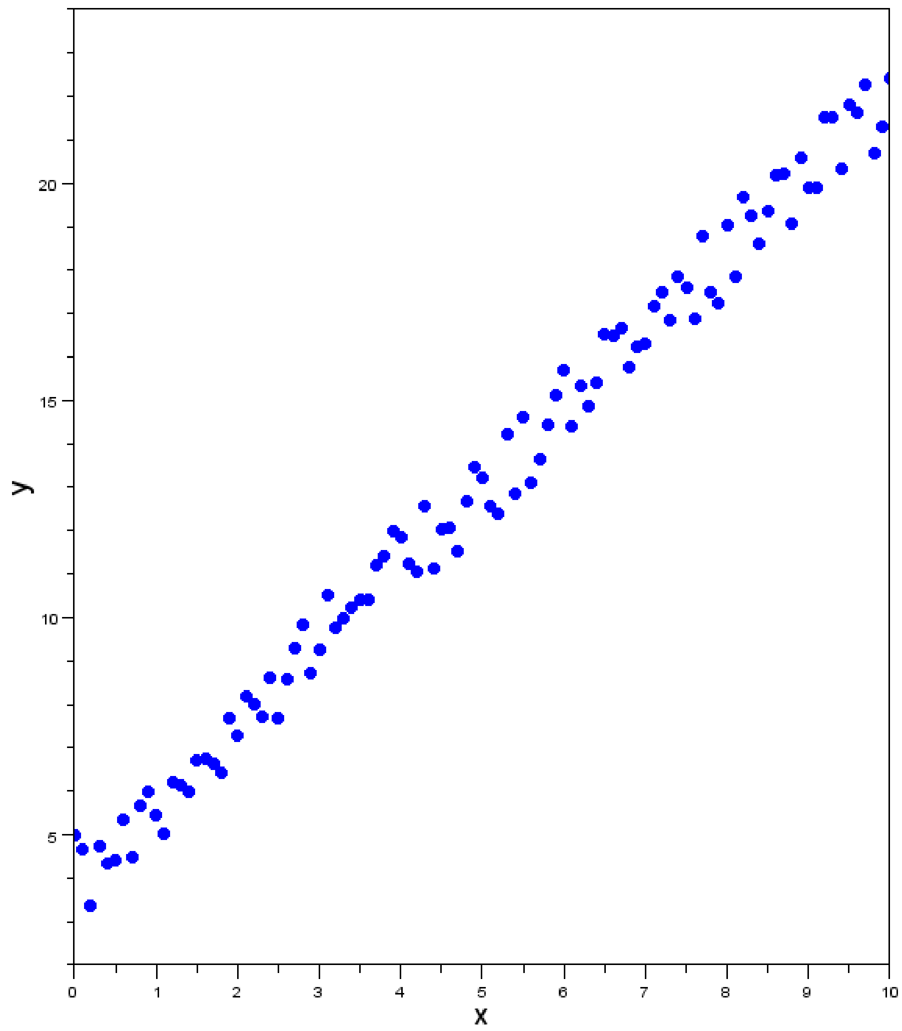
```
x = [0:0.1:10]';
n = 10*rand(length(x),1);

x0 = 1.0*x + 0.0*n;
y = 2*x0 + 3;
plot(x, y, 'b.');

s2x = mean(x.^2) - mean(x)^2;
s2y = mean(y.^2) - mean(y)^2;
Cov = mean(x.*y) - mean(x)*mean(y)
p2 = Cov / sqrt(s2x*s2y)

Cov =   17.0000

p2 =    1.0000
```



90% data, 10% noise

- $\rho^2 = 0.9943$

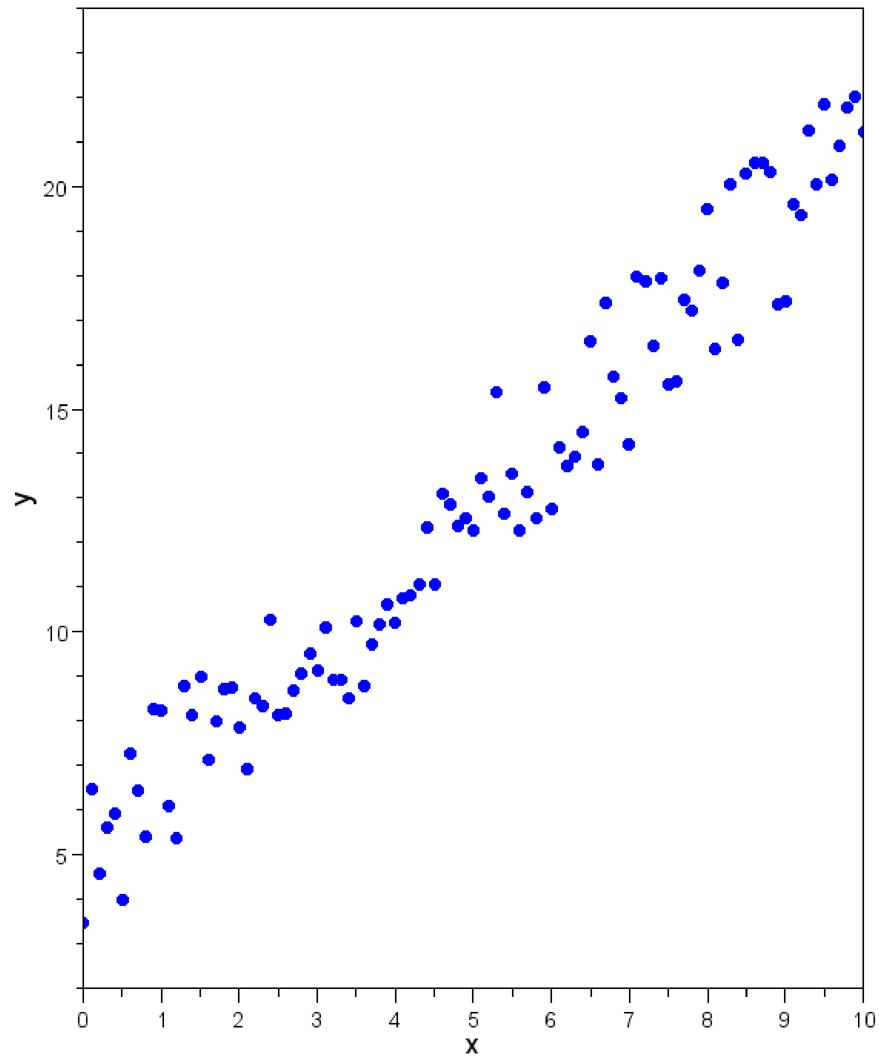
```
x = [0:0.1:10]';
n = 10*rand(size(x0));

P = zeros(100,1);

x0 = 0.9*x + 0.1*n;

y = 2*x0 + 3;
s2x = mean(x.^2) - mean(x)^2;
s2y = mean(y.^2) - mean(y)^2;
Cov = mean(x.*y) - mean(x)*mean(y)
p2 = Cov / sqrt(s2x*s2y)

Cov =    15.5010

p2 =     0.9943
```



80% data / 20% noise

- $\rho^2 = 0.9700$

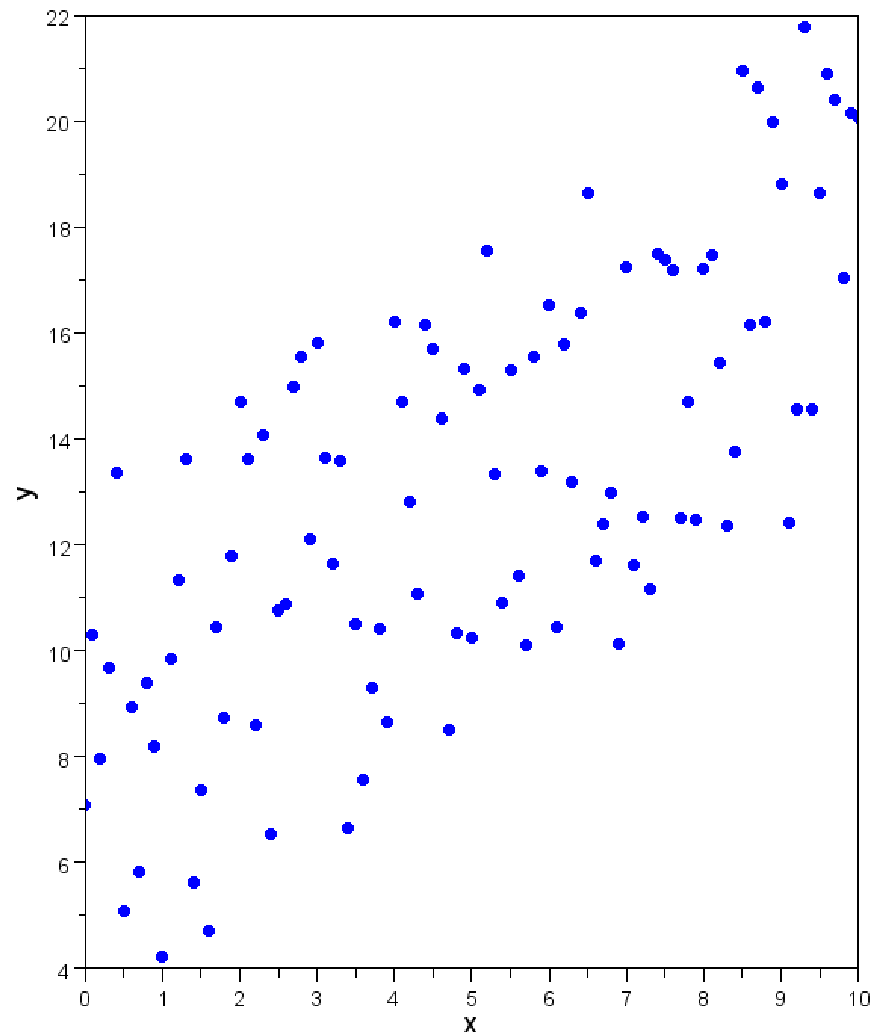
```
x = [0:0.1:10]';
n = 10*rand(size(x0));

x0 = 0.8*x + 0.2*n;

y = 2*x0 + 3;
s2x = mean(x.^2) - mean(x)^2;
s2y = mean(y.^2) - mean(y)^2;
Cov = mean(x.*y) - mean(x)*mean(y)
p2 = Cov / sqrt(s2x*s2y)

plot(x,y,'.')

Cov =   13.8326

p2 =    0.9700
```



50% data / 50% noise

- $\rho^2 = 0.7074$

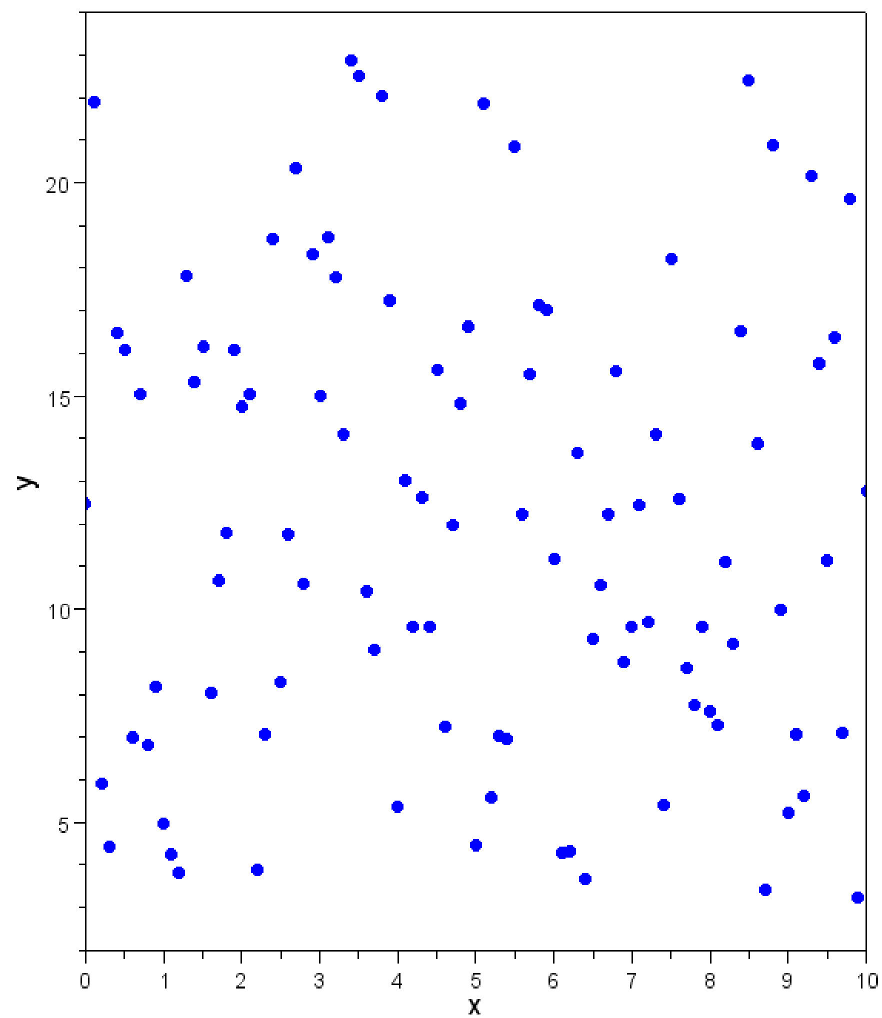
```
x = [0:0.1:10]';
n = 10*rand(size(x0));

x0 = 0.5*x + 0.5*n;

y = 2*x0 + 3;
s2x = mean(x.^2) - mean(x)^2;
s2y = mean(y.^2) - mean(y)^2;
Cov = mean(x.*y) - mean(x)*mean(y)
p2 = Cov / sqrt(s2x*s2y)

plot(x,y,'.')

Cov =     8.3611

p2 =    0.7074
```



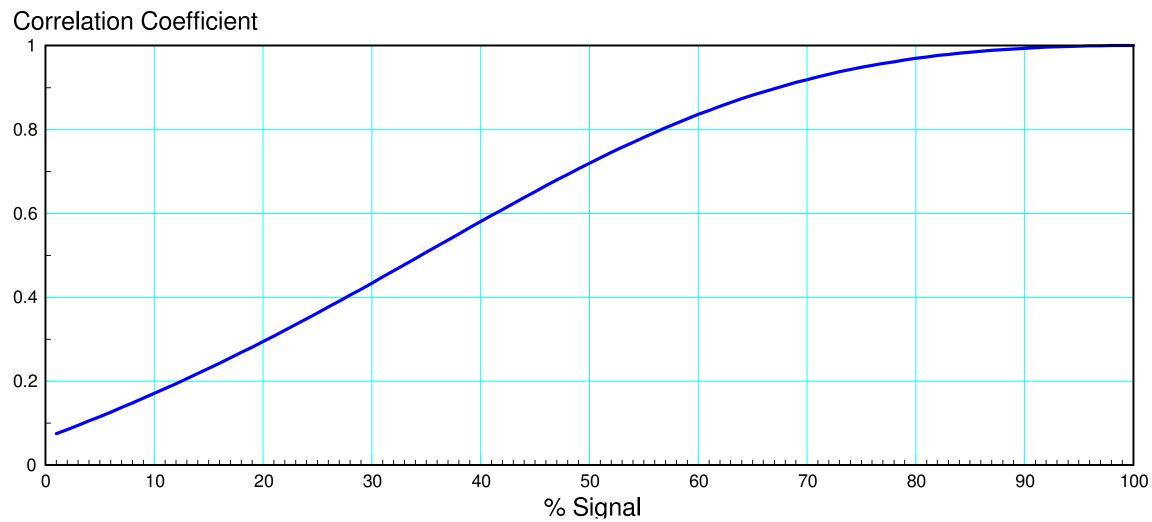
0% Data, 100% Noise

- $\rho^2 = 0.2429$

```
x = [0:0.1:10]';
n = 10*rand(size(x0));

x0 = 0.0*x + 1.0*n;

y = 2*x0 + 3;
s2x = mean(x.^2) - mean(x)^2;
s2y = mean(y.^2) - mean(y)^2;
Cov = mean(x.*y) - mean(x)*mean(y)
p2 = Cov / sqrt(s2x*s2y)

plot(x,y,'.')

Cov =     4.0005

p2 =   0.2494
```



What's the relationship between the correlation coefficient and the contribution of the signal to your measurement?




```
x = [0:0.1:10]';
n = 10*rand(size(x));

P = zeros(100,1);

for i=1:100
   a = i/100;
   x0 = a*x + (1-a)*n;

   y = 2*x0 + 3;
   s2x = mean(x.^2) -
mean(x)^2;
   s2y = mean(y.^2) - mean(y)^2;
   Cov = mean(x.*y) - mean(x)*mean(y)
   p2 = Cov / sqrt(s2x*s2y)
   P(i) = p2;
   end
```

Moral: Don't be impressed with correlation coefficients less than 0.8