C.Candan
EE230-METU

On Correlation Coefficient

The correlation coefficient indicates the degree of "linear dependence" of two random variables. It is defined as

$$r_{xy} = \frac{E\{(x-\overline{x})(y-\overline{y})\}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

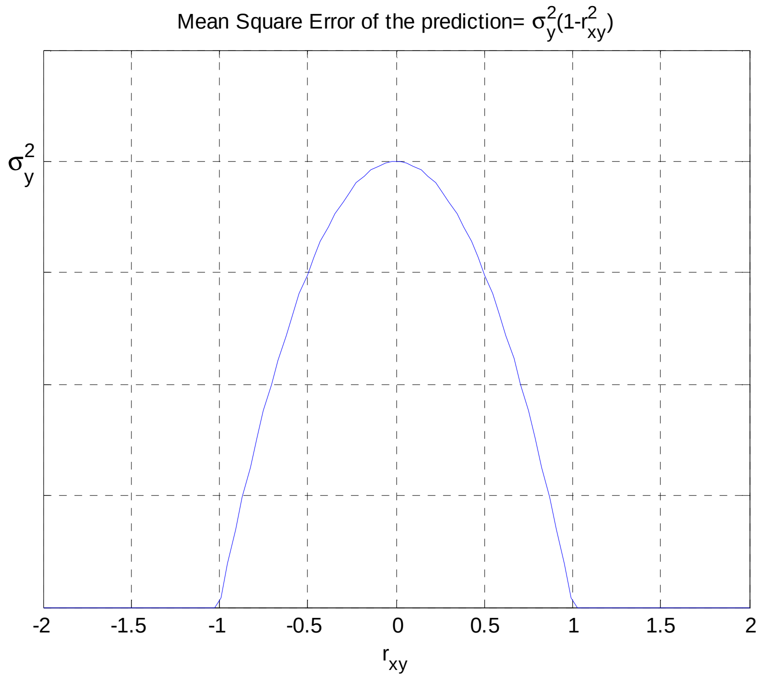
<u>Properties:</u>

1. $|r_{xy}| \leq 1$
2. If $|r_{xy}| = 0$ then $\underset{\sim}{x}$ and $\underset{\sim}{y}$ are called uncorrelated random variables. (Note that two independent variables are guaranteed to be uncorrelated; but the reverse is not true in general. So there can be two random variables which are uncorrelated, but dependent.)
3. $|r_{xy}| = 1 \Leftrightarrow y = ax + b$ Here a and b are non-random parameters, i.e. scalars. This relation shows that when $|r_{xy}| = 1$, then the random variable $\underset{\sim}{y}$ is a linearly related to $\underset{\sim}{x}$ and vice-versa. If $|r_{xy}| = 1$, knowing $\underset{\sim}{y}$ or $\underset{\sim}{x}$ is sufficient to determine the other one through $y = ax + b$. So knowing one of two random variables is as good as knowing the both of them.
4. In many applications, we can estimate the correlation coefficient between two random variables by conducting experiments. In practice we use the correlation coefficient to predict the value of $\underset{\sim}{y}$ (something of interest) when we can only observe $\underset{\sim}{x}$. We are not lucky to observe $\underset{\sim}{y}$ directly in many applications. If $\underset{\sim}{y}$ and $\underset{\sim}{x}$ are closely related, then we may expect that we can reliably predict $\underset{\sim}{y}$ from $\underset{\sim}{x}$.

   Lets say we are interested in $\underset{\sim}{y}$; but have only $\underset{\sim}{x}$ and we know the correlation coefficient between $\underset{\sim}{x}$ and $\underset{\sim}{y}$. You will learn in some other courses that we can predict $\underset{\sim}{y}$ as follows $\hat{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x-\overline{x}) + \overline{y}$. This is the best linear prediction of $\underset{\sim}{y}$ in the mean square sense. (You will also hear about *mean square sense* at these courses.)

Remember that we have noted in item 3 the following: If $|r_{xy}|=1$, the knowing $\underset{\sim}{x}$ or $\underset{\sim}{y}$ is as good as knowing both of them. Therefore we expect to have zero prediction error in this case. For other r_{xy} values, the value of the prediction error is not immediately clear.

The graph given below shows the mean square error (approximation error) for a general value of r_{xy}. As expected, the mean square error is zero, when $|r_{xy}|=1$ and as the magnitude of correlation coefficient decreases, the error increases. The error reaches its maximum when two random variables are uncorrelated.



```
rxy=linspace(-2,2,100).*rect(linspace(-2,2,100),-1,1);
plot(linspace(-2,2,100),2*(1-rxy.^2).*rect(linspace(-2,2,100),-1,1));
grid on; xlabel('r_{xy}');
title('Mean Square Error of the prediction= \sigma_y^2(1-r^2_{xy})');
axis([-2 2 0 2.5])
```

[For more info Hayes, Statistical Digital Signal Processing and Modeling, p. 70]

<u>Examples with Scatter Plots:</u>

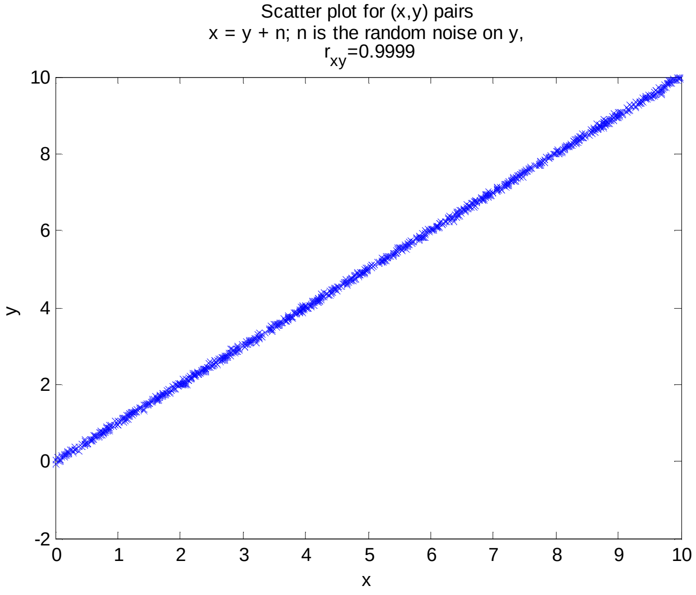
Lets say that we want to learn $\underset{\sim}{y}$; but we can only observe $\underset{\sim}{x}$. Let the observation model be given as

$$\underset{\sim}{y} = \underset{\sim}{x} + \underset{\sim}{n}$$

Here $\underset{\sim}{n}$ is the effect of noise. (You can assume zero mean noise without any harm or loss of generality.) The correlation coefficient between $\underset{\sim}{x}$ and $\underset{\sim}{y}$ can be calculated as

$$r_{xy} = \sqrt{\frac{\sigma_x^2}{\sigma_x^2 + \sigma_n^2}}\,.$$

Lets start with the case of little noise When noise is little, i.e. variance of noise is small; r_{xy} is close to 1.

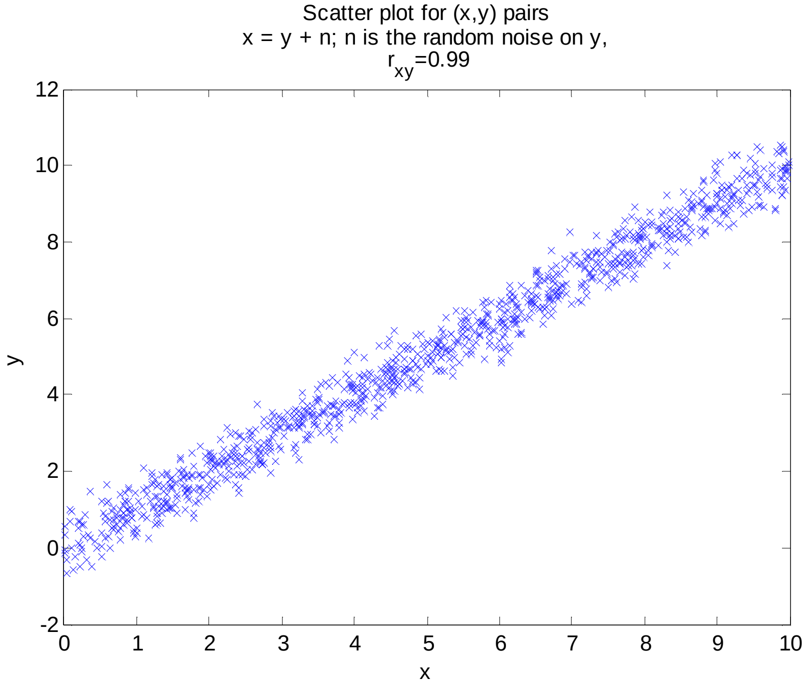
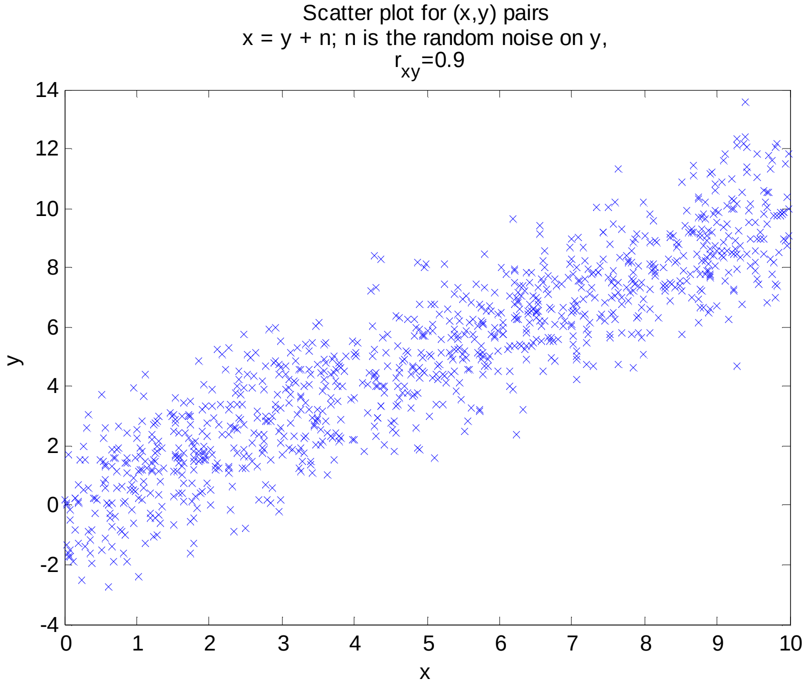
$r_{xy} = 0.999$	

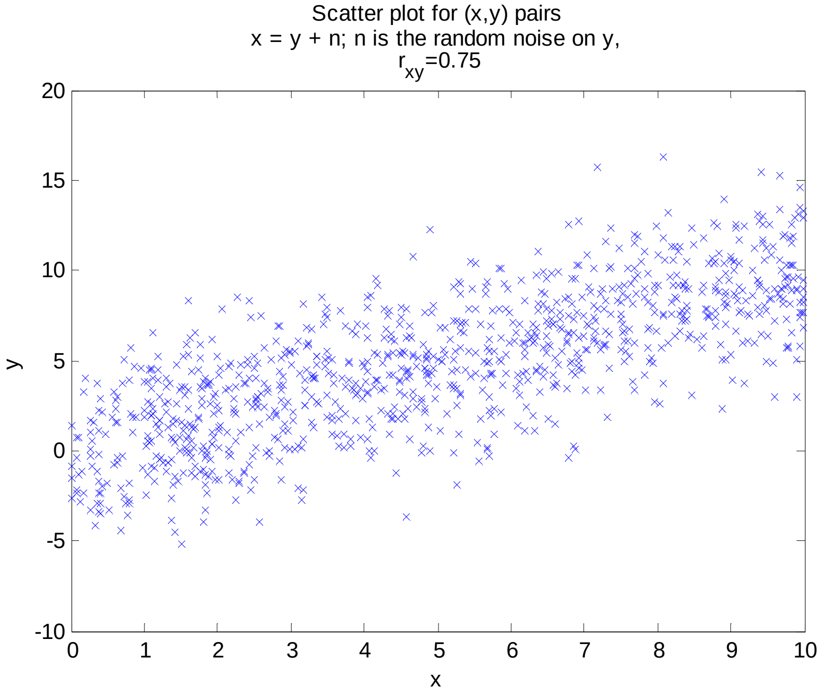
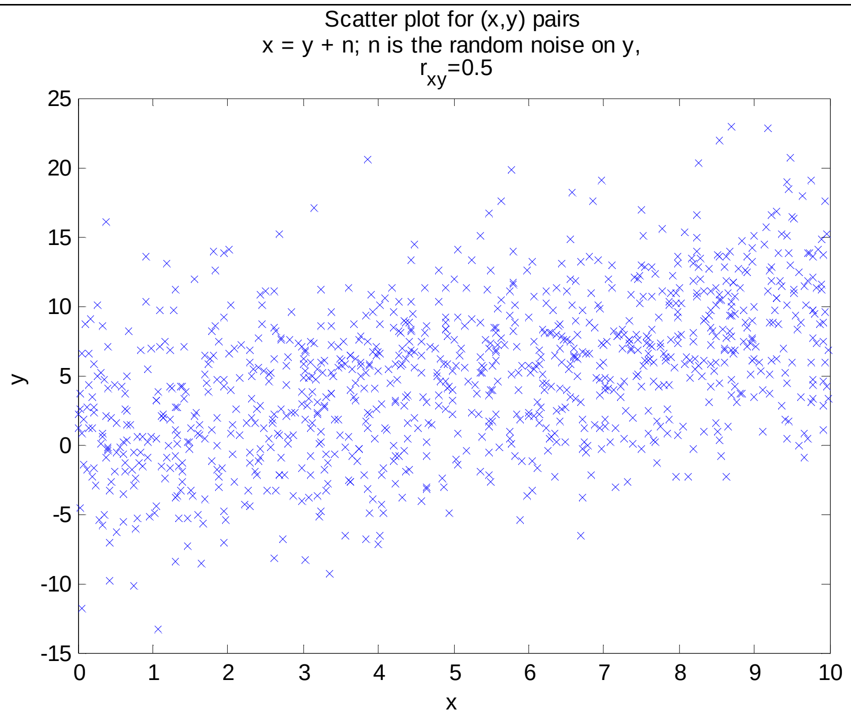
The plot given above is called scatter plot and it is drawn by randomly generating $\underset{\sim}{x}$ and $\underset{\sim}{n}$ and calculating $\underset{\sim}{y}$ through $\underset{\sim}{y} = \underset{\sim}{x} + \underset{\sim}{n}$. If there were no noise, y = x ; but unfortunately there is noise in any observation.

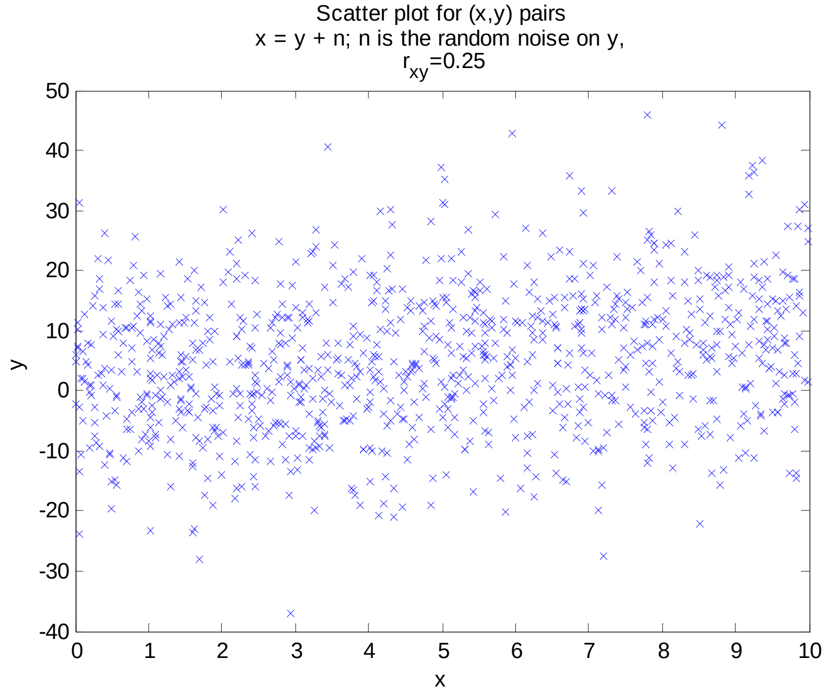
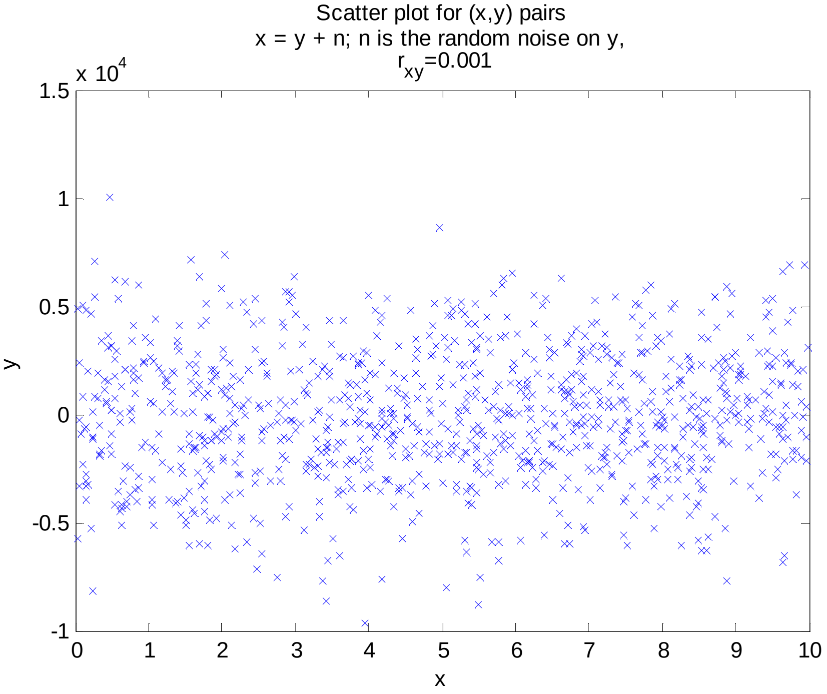
The scatter plot is drawn by putting cross marks (x) where the randomly generated $\underset{\sim}{x}$ and calculated $\underset{\sim}{y}$ are on the (x,y) plane. There are 1000 crosses in the given figure.

So we conclude from this figure, when there is little noise, knowing $\underset{\sim}{x}$ can be as good as knowing $\underset{\sim}{y}$, which is wonderful.

Below we have some other scatter plots. The noise level is higher in these plots, therefore there is a bigger spread around the y=x line.

$r_{xy} = 0.99$	
$r_{xy} = 0.9$	

$r_{xy} = 0.75$	
$r_{xy} = 0.5$	

$r_{xy} = 0.25$	
$r_{xy} = 0.01$	

So as a conclusion, the correlation coefficients show how much two random variables are related to each other in a linear way.

Matlab code for scatter plots:

```
x=10*rand(1,1000);

rxy=0.25;

sigmax2=10^2/12;
%rxy= sqrt(sigmax2/(sigmax2+sigman2)
sigman2 = (1/rxy^2-1)*sigmax2,

y=x + sqrt(sigman2)*randn(size(x));
plot(x,y,'x');
title(['Scatter plot for (x,y) pairs' char(10) 'x = y + n; n is the
random noise on y,' char(10) ' r_{xy}=' num2str(rxy)]);
xlabel('x'),ylabel('y');
```