C.Candan
EE230/503-METU

On Correlation Coefficient

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

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

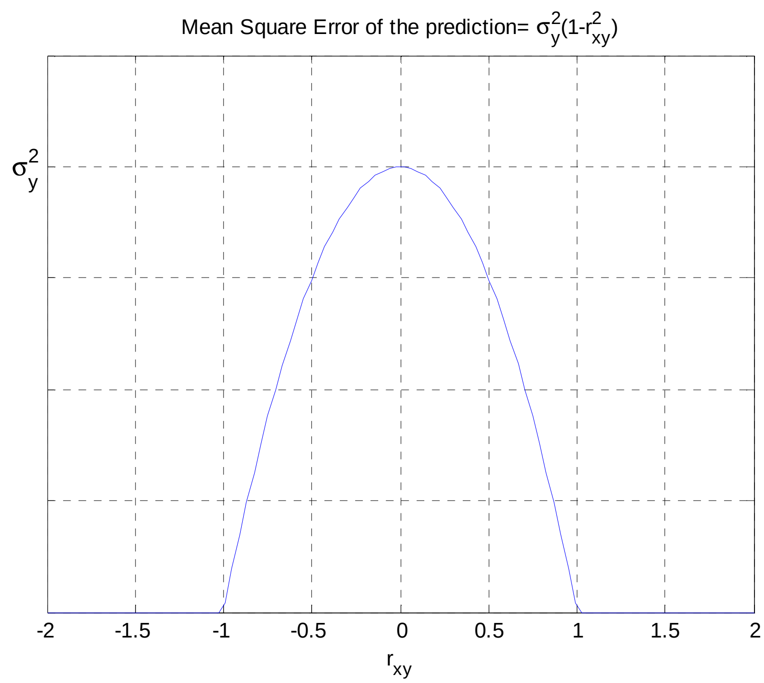
Properties:

1. $|r_{xy}| \leq 1$. (See appendix for the proof of this property.)
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. (See appendix for the proof of this property.)
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 $\hat{y} = r_{xy} \dfrac{\sigma_y}{\sigma_x}(x - \bar{x}) + \bar{y}$. 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>

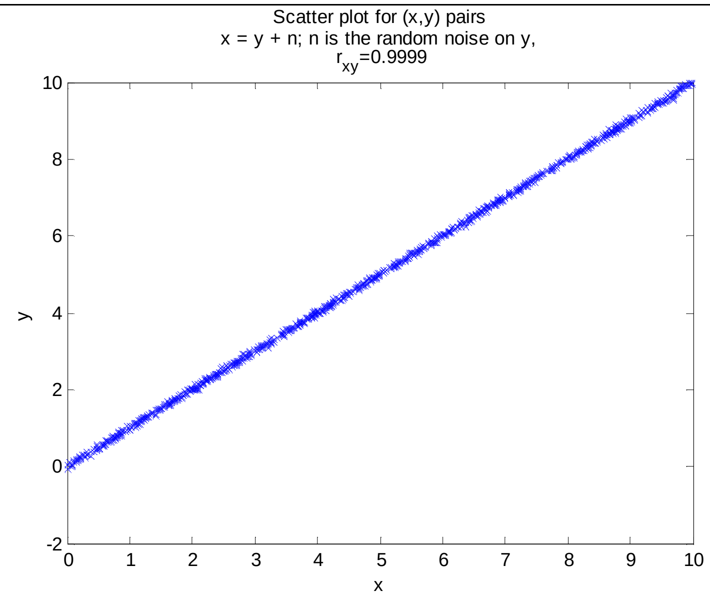
Lets say that we want to learn $\underset{\sim}{y}$; but we can only observe $\underset{\sim}{x}$. Let the observation model be given as

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

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.

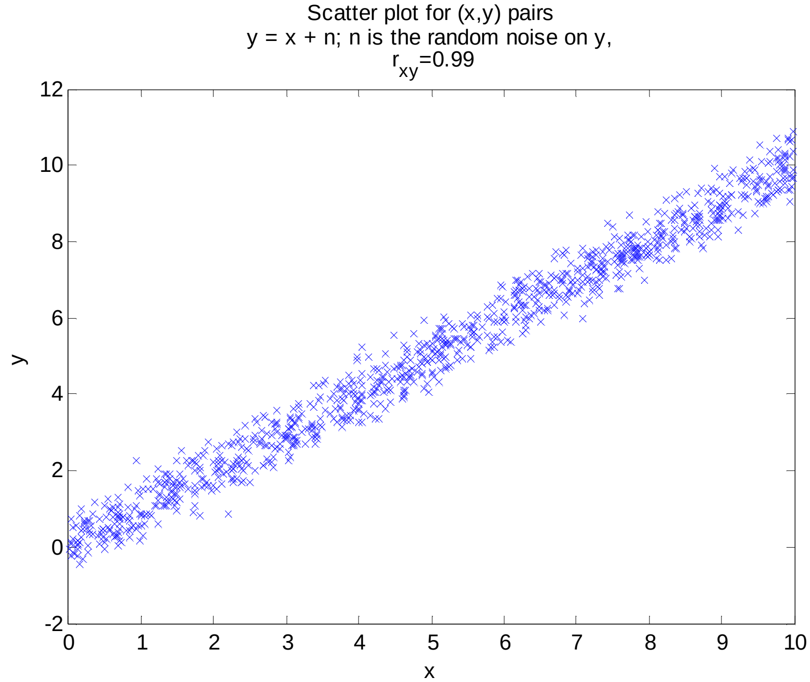
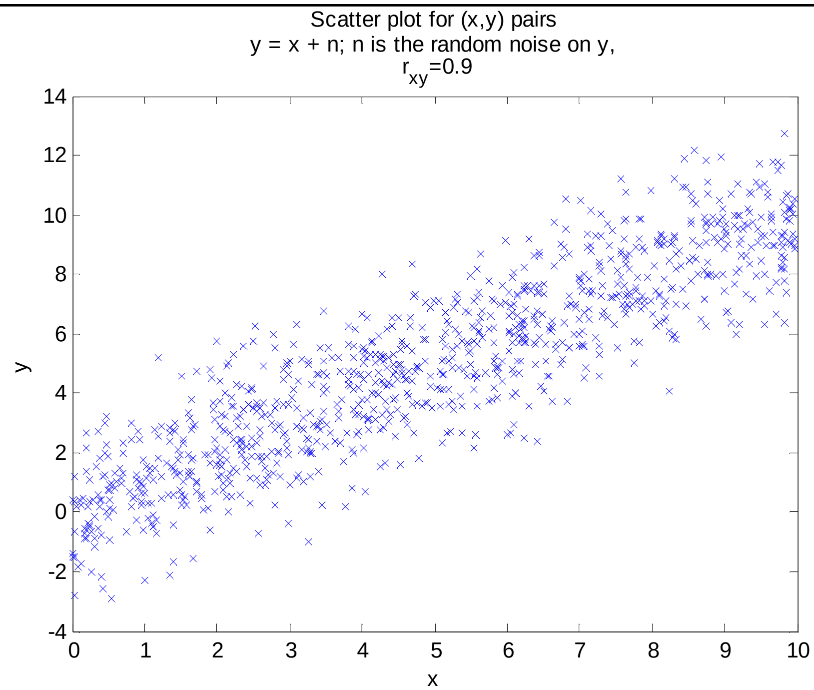
$r_{xy} = 0.999$	

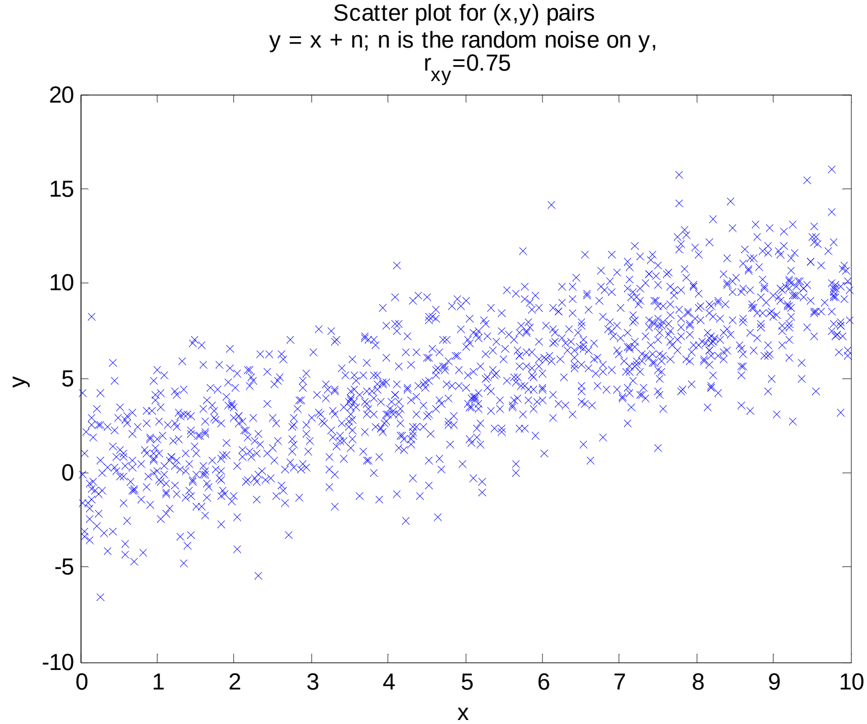
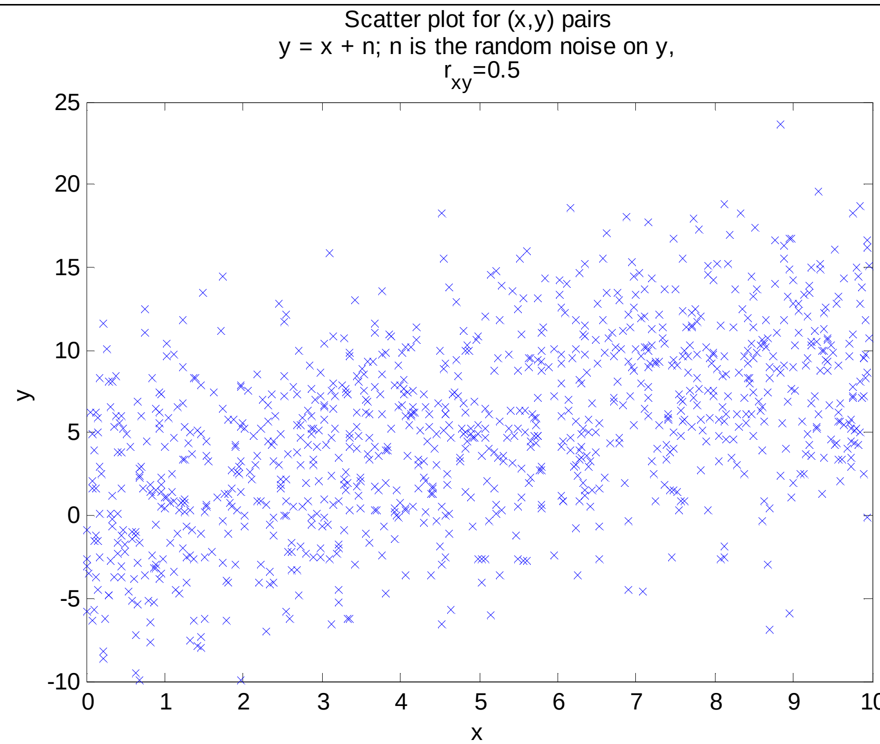
The plot given above is called scatter plot and it is drawn by randomly generating $\underset{\sim}{x}$ and $\underset{\sim}{n}$ and calculating $\underset{\sim}{y}$ through $\underset{\sim}{y} = \underset{\sim}{x} + \underset{\sim}{n}$. If there were no noise, y = x ; but unfortunately there is noise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

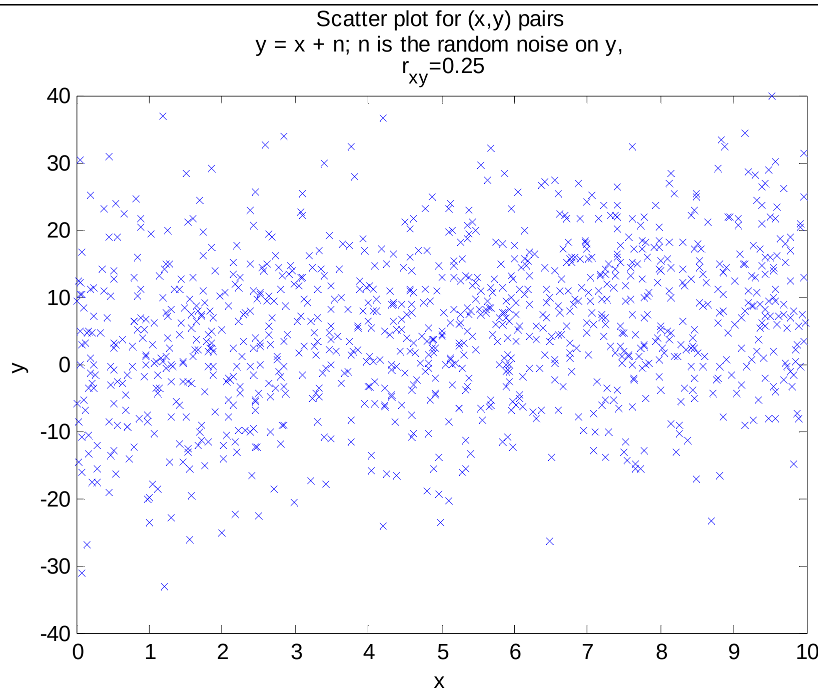
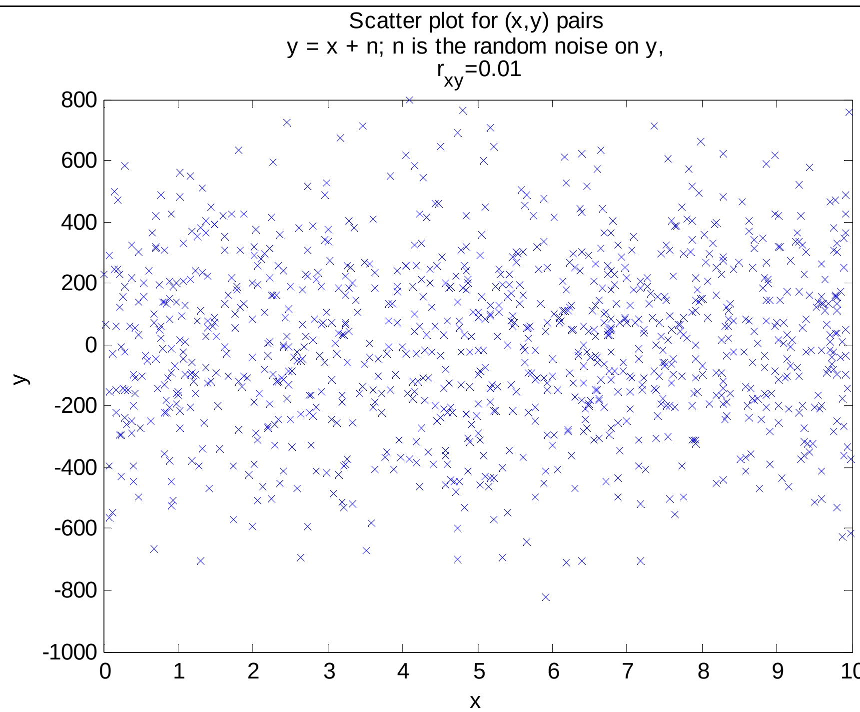
The scatter plot is drawn by putting cross marks (x) where the randomly generated $\underset{\sim}{x}$ and calculated $\underset{\sim}{y}$ are on the (x,y) plane. There are 1000 crosses in the given figure.

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');
```

Non-Linearly Related Random-Variables and Correlation Coefficient

In the previous section, we have tried to interpret the correlation coefficient for a linear observation model. Linear observation model means that the signal of interest is mapped to the output through a linear function. In the example presented in the previous section, the model is extremely simple (but useful) one, $y = x + n$. In this section, elaborate further on the same topic; but we switch to the non-linear observation models such as $y = x^2 + n$.

As in the previous example, lets assume that x is uniformly distributed in [0,Δ]. Then $E\{x\} = \dfrac{\Delta}{2}$, $E\{x^2\} = \dfrac{\Delta^2}{3}$, $E\{x^3\} = \dfrac{\Delta^3}{4}$ and so on. We will construct a non-linear function of x in the form $y = f(x) = ax^2 - bx$ such that the correlation coefficient of x and y is zero! (Note that, we are not adding any noise to the observations. The correlation is zero in the absence of noise!)

The correlation coefficient is expressed as follows:

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

It is clear that for the correlation coefficient to be zero, $E\{xy\} - \overline{xy} = 0$.

Let's calculate $E\{xy\}$:

$$\begin{aligned} E\{xy\} &= E\{x(ax^2 - bx)\} \\ &= aE\{x^3\} - bE\{x^2\} \\ &= a\frac{\Delta^3}{4} - b\frac{\Delta^2}{3} \end{aligned}$$

Similarly we can calculate $\overline{xy}$ as follows: $\overline{xy} = \frac{\Delta}{2}\left(a\frac{\Delta^2}{3} - b\frac{\Delta}{2}\right) = a\frac{\Delta^3}{6} - b\frac{\Delta^2}{4}$. Now we are ready to evaluate $E\{xy\} - \overline{xy}$:

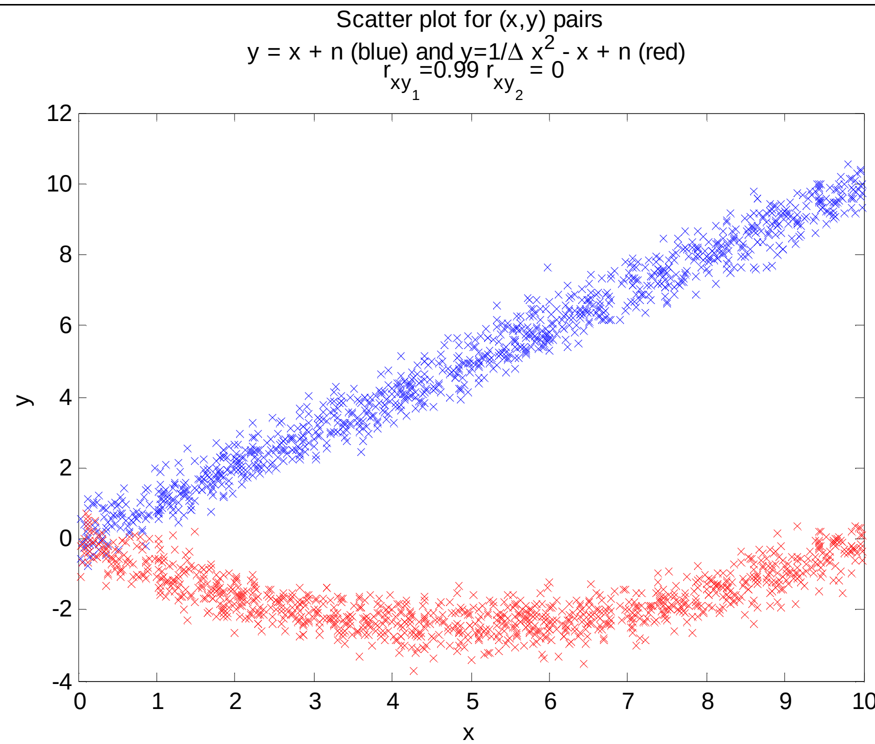
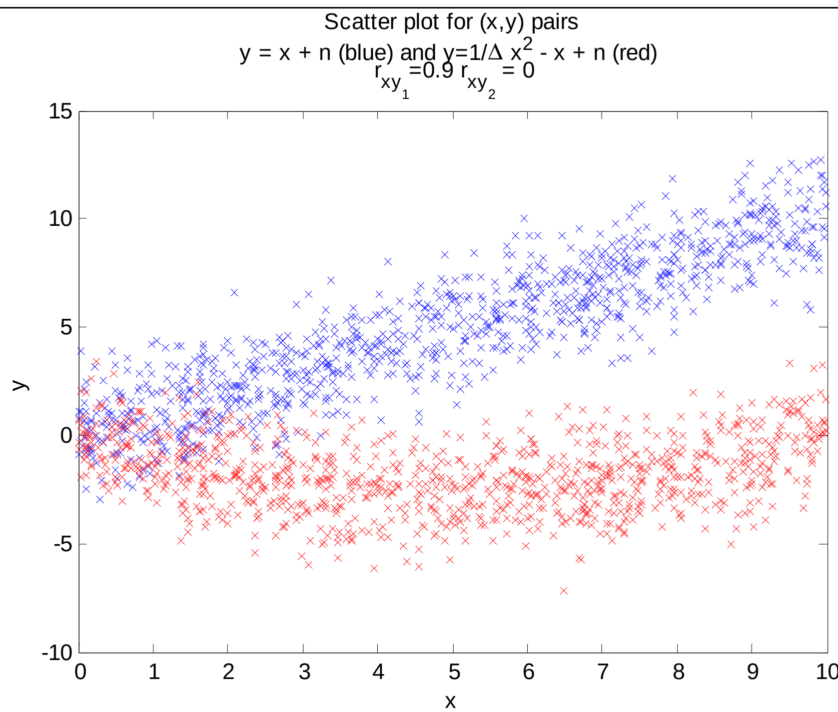
$$E\{xy\} - \overline{xy} = a\left(\frac{\Delta^3}{4} - \frac{\Delta^3}{6}\right) - b\left(\frac{\Delta^2}{3} - \frac{\Delta^2}{4}\right) = \frac{\Delta^2}{12}(a\Delta - b)$$

From the last relation, we can conclude that when $\frac{a}{b} = \frac{1}{\Delta}$, the correlation coefficient of *x* and *y* is zero.

The following figure presents, the scatter plot for the linear observation model and the non-linear observation model. We are assuming that both observation models have the same noise, i.e. noise is zero-mean Gaussian and have identical variance.

For the plot given in top part of the figure, the correlation coefficient for linear model is set to 0.99. For the non-linear observation model, it is equal to 0 since we have set $a = \frac{1}{\Delta}$ and $b = 1$.

From these figures, we can see the correlation coefficient of two random variables having a non-linear relation between them should be treated with care. From these figures, it is clear that when the effect of noise is little, it is possible to say something about *x* given observation *y* for both models. At least, it is possible to reduce the set of possibilities for the unknown *x* given y. Unfortunately, the correlation coefficient of the non-linear observation model is equal to zero irrespective of the noise level corrupting the observations. Hence correlation coefficient and related ideas are especially useful for linear observation models.

$r_{xy_1} = 0.99$ $r_{xy_2} = 0$	
$r_{xy_1} = 0.9$ $r_{xy_2} = 0$	

The following is the Matlab code generating the figure presented above.

```
Delta=10;MCnum=1e3;
x=Delta*rand(1,MCnum);

rxy=0.9;

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

y1=x + sqrt(sigman2)*randn(size(x));
y2=1/Delta*x.^2 - x + sqrt(sigman2)*randn(size(x));
plot(x,y1,'x'); hold all;
plot(x,y2,'xr'); hold off;
title(['Scatter plot for (x,y) pairs' char(10) 'y = x + n (blue)' ...
       ' and y=1/\Delta x^2 - x + n (red) ' char(10) ...
       ' r_{xy_1}=' num2str(rxy)...
       ' r_{xy_2} = 0' ]);
xlabel('x'),ylabel('y');

corr_coef1 = 1/MCnum*sum((x-mean(x)).*(y1-mean(y1)))/sqrt(var(x)*var(y1)),
corr_coef2 = 1/MCnum*sum((x-mean(x)).*(y2-mean(y2)))/sqrt(var(x)*var(y2)),
```

The last two lines of the Matlab code generates an estimate for the correlation coefficient. The estimate is produced by estimating the mean, variance and cross-correlation of random variables from the experimental data.

When MCnum is set to 10,000 and the script is run, we get the following result:

corr_coef1 =

0.9001

corr_coef2 =

-0.0085

This result shows that the correlation coefficient for the linear model is almost equal to 0.9 (as expected) and there is indeed very little correlation between x and y for the non-linear model.

Appendix:

Proof of $|r_{xy}|=1 \Leftrightarrow y=ax+b$

It can be noted that that from the definition of r_{xy} that r_{xy} does not depend on the mean values of x and y. Without loss of any generality, we assume that x and y are zero mean. Then the result to be proved and the definition of r_{xy} reduces to $|r_{xy}|=1 \Leftrightarrow y=ax$, $r_{xy}=\frac{E\{xy\}}{\sqrt{E\{x^2\}E\{y^2\}}}$ respectively, for zero-mean random variables x and y.

i. **Proof of** $|r_{xy}|=1 \Leftarrow y=ax$

If $y=ax$, then $E\{xy\}=a\sigma_x^2$, $E\{y^2\}=a^2\sigma_x^2$, and $E\{x^2\}=\sigma_x^2$, then $r_{xy}=\frac{E\{xy\}}{\sqrt{E\{x^2\}E\{y^2\}}}=\frac{a\sigma_x^2}{\sqrt{a^2\sigma_x^2\sigma_x^2}}=1.$

ii. **Proof of** $|r_{xy}|=1 \Rightarrow y=ax$

Let $P(\psi)$ be a quadratic polynomial in ψ. $P(\psi)$ is defined as follows: $P(\psi)=E\{(x\psi-y)^2\}=E\{x^2\}\psi^2-2E\{xy\}\psi+E\{y^2\}=a\psi^2+b\psi+c$. It is clear that $P(\psi)\geq 0$ for all ψ values. Then the discriminant of $P(\psi)$, i.e. $\Delta=b^2-4ac$, should be either 0 or negative valued. The discriminant can be calculated as $\Delta=(2E\{xy\})^2-4E\{x^2\}E\{y^2\}$. Since $\Delta\leq 0$, $(E\{xy\})^2\leq E\{x^2\}E\{y^2\}$ and then $|r_{xy}|=\left|\frac{E\{xy\}}{\sqrt{E\{x^2\}E\{y^2\}}}\right|\leq 1$, which is the first property.

If $|r_{xy}|=1$, then $\Delta=(2E\{xy\})^2-4E\{x^2\}E\{y^2\}=0$. Therefore there is a specific ψ value called ψ_x for which $P(\psi_x)=0$. This leads to $P(\psi_x)=E\{(x\psi_x-y)^2\}=0$. The last relation shows that $y=\psi_x x$.