Consider two random variables X and Y with

$$\mu_X = E(X), \sigma_X^2 = Var(X), \mu_Y = E(Y), \sigma_Y^2 = Var(Y).$$

In order to study the "relationship" between the two random variables, we need a numerical measure that describes the relationship. The **covariance** between two random variables which are observed in pairs is such measure, and is defined as follows.

$$Cov(X,Y) = \sigma_{XY} = E\{(X-\mu_X)(Y-\mu_Y)\} = E(XY) - \mu_X\mu_Y$$

Notet that $\sigma_{XY} > 0$ implies, on average, $(X-\mu_X)(Y-\mu_Y) > 0$, that is, that is, values of *X* larger (smaller) than its mean μ_X tend to be associated with values of *Y* larger (smaller) than its mean μ_Y. Also $\sigma_{XY} < 0$ implies, on average, $(X-\mu_X)(Y-\mu_Y) < 0$, that is, values of *X* larger (smaller) than its mean μ_X tend to be associated with values of *Y* smaller (larger) than its mean μ_Y. Therefore, a positive covariance implies a positive relation between two random variables, and a negative covariance a negative relation.

The sample covariance is

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}\right)$$

when a random sample consists of n pairs of observations (X_i, Y_i) for *i*=1, 2, ..., n.

Example 1. Find the sample covariance between the square footage in thousands (*X*) and the annual sales in millions of dollars (*Y*).

i	x_i	y_i	x_iy_i
1	1.7	3.7	6.29
2	1.6	3.9	6.24
3	2.8	6.7	18.76
4	5.6	9.5	53.2
5	1.3	3.4	4.42
6	2.2	5.6	12.32
7	1.3	3.7	4.81
8	1.1	2.7	2.97
9	3.2	5.5	17.6
10	1.5	2.9	4.35
11	5.2	10.7	55.64
12	4.6	7.6	34.96
13	5.8	11.8	68.44
14	3	4.1	12.3
	40.9	81.8	302.3

HC Homework: Do the following problem.. (You may use MS Excel.)

Problem 1. Using the data in the above example, obtain the covariance between the square footage and the annual sales **in Euros** and compare it with the covariance obtained in Example 1.

To compute the sample covariance using MS Excel, do
Tools>Data Analysis>Covariance, and enter the range of *x*- and *y*- variables.
For the data in Example 1 we get

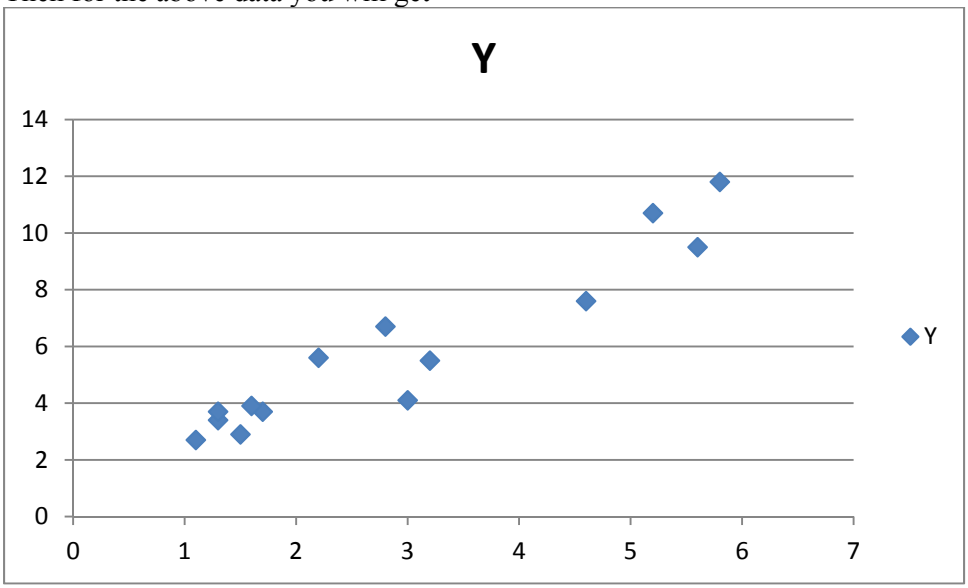
	X	*Y*
X	2.708827	
Y	4.523367	8.353878

Note: Unfortunately, this version of MS Excel used *n* (sample size) as denominator instead of $n-1$. Therefore, we need to multiply the numbers by *n* (in this example 14) and divide by $n-1$ in order to get the "correct" covariance and variances:

	X	*Y*
X	2.917198	
Y	4.871319	8.996484

There are two Excel functions COVARIANCE.**P** and COVARIANCE.**S** for the population and the sample covariance, respectively.

To make a scatter plot of the above data, highlight the data and Chart Wizard>XY (Scatter). Then for the above data you will get



As you will see in Problem 1, the magnitude of covariance depends on the units of variables quoted. Therefore, the magnitude of variable cannot effectively represent the "strength" of the relationship.

The **correlation coefficient** between two random variables X and Y is

$$Corr(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = Cov\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right)$$

It turns out the correlation coefficient is the covariance between the standardized variable of X and the standardized variable of Y. Since the standardized variables are unit free, so is the correlation coefficient. The sample correlation coefficient is

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

Example 2. Using the data in Example 1, find the sample correlation coefficient.

To compute the sample covariance using MS Excel do
Tools>Data Analysis>Correlation, and enter the range of x- and y- variables.
For the above data we get

	X	*Y*
X	1	
Y	0.950883	1

You may use the Excel function **CORREL**.

HC Homework: 3.46 on p. 137 and do the following problem.. (You may use MS Excel.)
Problem 2. Using the data in Example 1, obtain the correlation between the square footage and the annual sales **in Euros** and compare it with the correlation obtained in Example 2.

MSL Homework: 3.44 on p. 136 (You may use MS Excel.)

Properties of the correlation coefficient.

1. $-1 \le \rho \le 1$
2. $\rho = 1$ represents a perfect positive linear association.
3 $\rho = -1$ represents a perfect negative linear association.
4. $|\rho|$ closer to 1 represents stronger linear association.
5 $\rho = 0$ represents no linear association, but the variables can have some relation such as quadratic relation.

MSL Homework: 13.1, 13.2, 13.3

When two variable are correlated, we are often interested in finding the "precise" relationship using a mathematical model and in predicting one variable of main interest, which is called the dependent variable or response variable and denoted by Y, using the other variable, which is called the independent variable or predictor variable and denoted by X.

There are two types of relationship considered in this chapter.
Deterministic relationship: each value of X is paired with one and only one value of Y, and Y can be predicted with certainly for a given value of X.

Stochastic (**Statistical**) relationship: each value of *X* is associated with a whole probability distribution of values of *Y*, and Y cannot be predicted with certainly for a given value of *X*, but the knowledge of a value of *X* helps in predicting *Y*.

A simple and popular mathematical model to describe a stochastic relationship is the following
Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where *Y* is the dependent variable and *X* is the independent variable, and the random variable ε is called the error satisfying $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$.

The purposes of regression analysis include

1) to better understand the relationship between the dependent variable and the independent variable(s) through mathematical models; and
2) to predict the values of the dependent variable with given values of the independent variables.

The assumptions about the ε in turn yield $E(Y) = \beta_0 + \beta_1 X$, that is, the mean of *Y* is a linear function of *X* and thus the knowledge about the value of *X* is useful in predicting the mean value of *Y* (as well as individual value of *Y*). The line equation

$$E(Y) = \beta_0 + \beta_1 X$$

is called the **regression line** (and in general **regression function**). They also yield $Var(Y) = \sigma^2$, which means the variability of *Y* is constant regardless of the level of *X*.
Fore regression analysis we have a random sample of n pairs of observations (X_i, Y_i) for *i*=1,...,n and for these observations we have

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where the ε_i are independent.

In the above simple linear regression model we have three (unknown) parameters that need to be estimated. They are β_0 and β_1, which are called the model parameters, and σ^2, which is the variance of the error. Estimation of these parameters are done by the method of least squares, which, roughly speaking, finds a fitted line that goes through the points on a scatter plot such that the line is as "close" as possible to all the points[1]. The least squares estimators of β_0 and β_1, denoted by b_0 and b_1, respectively are

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{and} \quad b_0 = \bar{Y} - b_1 \bar{X}.$$

Note that $b_1 = \frac{S_{xy}}{S_x^2} = r_{xy}\frac{S_y}{S_x}$.

If we replace the unknown parameters with their estimates, we obtain the **estimated regression line**, also called, **fitted line**

[1] Techincally b_0 and b_1 are chosen to minimize $\sum_{i=1}^{n}\{Y_i - (b_0 + b_1 X_i)\}^2$, the residual sum of squares.

$$\hat{Y} = b_0 + b_1 X .$$

MSL Homework: 13.5
HC Homework: 13.9 (Use Excel/PHStat)

For the *i*-th observation (X_i, Y_i), we can obtain the corresponding fitted value $\hat{Y} = b_0 + b_1 X_i$. The difference between the observed value Y_i and the fitted value $\hat{Y}_i$, that is, $Y_i - \hat{Y}_i$ is called the **residual**

$$e_i = Y_i - \hat{Y}_i .$$

One of the properties of the residuals is $\sum_{i=1}^{n} e_i = 0$.

The least squares estimator of σ^2, denoted by $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 ,$$

which is also called the **mean squared error**.

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 : \quad SST = SSReg + SSE$$

$$R^2 = \frac{SSReg}{SST} = 1 - \frac{SSE}{SST}$$

This measures the proportion of the total variation in *Y* (*SST*) explained by the regression model(*SSReg*), and is an overall measure of "goodness of fit."

For regression analysis using MS Excel do
Tools>Data Analysis>Regression, and enter the range of *y*- and *x*- variables. If you want confidence intervals other than 95%, check off Confidence Level and enter the confidence level. For the data in Example 1, we get an output on the next page.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950883
R Square	0.904179
Adjusted R Square	0.896194
Standard Error	0.96638
Observations	14

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	105.7476	105.7476	113.2335	1.823E-07
Residual	12	11.20668	0.93389		
Total	13	116.9543			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	
Intercept	0.964474	0.526193	1.832927	0.091727	-0.1820031	2.1109504	0.02664589	1
X	1.669862	0.156925	10.64112	1.82E-07	1.3279513	2.0117733	1.39017617	1

MSL Homework: 13.11, 13.12, 13.17
HC Homework: 13.21 (Use Excel/PHStat)

$(1-\alpha)100\%$ confidence interval for β_1

$$b_1 \pm t_{\alpha/2,n-2} s.e.(b_1)$$

$(1-\alpha)100\%$ confidence interval for β_0

$$b_0 \pm t_{\alpha/2,n-2} s.e.(b_0)$$

Testing hypothesis about the regression slope coefficient β_1

$H_0 : \beta_1 = \beta_{10}$ $H_1 : \beta_1 \neq \beta_{10}$	$H_0 : \beta_1 \geq \beta_{10}$ $H_1 : \beta_1 < \beta_{10}$	$H_0 : \beta_1 \leq \beta_{10}$ $H_1 : \beta_1 > \beta_{10}$

Test statistic: $T = \dfrac{b_1 - \beta_{10}}{s.e.(b_1)}$

Reject H_0, if

$\lvert t \rvert \geq t_{\alpha/2,n-2}$	$t \leq -t_{\alpha,n-2}$	$t \geq t_{\alpha,n-2}$

p-value

$P(\lvert T \rvert \geq \lvert t \rvert)$	$P(T \leq t)$	$P(T \geq t)$

Testing hypothesis about the regression intercept coefficient β_0

Replace β_1 with β_0 on the previous page.

Note the degrees of freedom in regression analysis is
the number of observation minus the number of parameters in the model.

MSL Homework: 13.40, 13.41, 13.43, 13.49
HC Homework: 13.47 (Use Excel/PHStat)

One of the purposes of regression analysis is to predict the mean value of *Y*, that is *E(Y)* given a value of *X*, say x_j. A point estimate of *E(Y)* given a value x_j is

$\hat{Y}_j = b_0 + b_1 X_j$ with standard error $\hat{\sigma}\sqrt{h_j}$, where $h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{(n-1)s_x^2}$ is called the leverage

Therefore the $(1-\alpha)100\%$ confidence interval for *E(Y)* is $\hat{Y}_j \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{h_j}$. Note $(n-1)s_x^2 = SSX$.

Another purpose of regression analysis is to predict an individual value of *Y* given a value of *X*, say, x_j. Then, a point estimate is ,again, $\hat{Y}_j = b_0 + b_1 X_j$ with standard error $\hat{\sigma}\sqrt{1+h_j}$. Therefore $(1-\alpha)100\%$ confidence interval for an individual value of *Y*, which is often called the **prediction interval**, is $\hat{Y}_j \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1+h_j}$.

Example 3. Suppose in Example 1, you want to estimate the mean annual sales of all stores with the size of 5 thousand square feet. Then it is 0.9645+1.6699×5=9.3140 (million dollars). Noting that the sample mean and standard deviation of X-variable are 2.9214 and 1.7080, respectively, (which are also computed from MS Excel) we calculate the standard deviation for $\hat{Y}_j$ as $0.9664\sqrt{0.185} = 0.4161$ noting that the leverage is $\frac{1}{14} + \frac{(5-2.9214)^2}{(14-1)\times(1.708)^2} = 0.1854$. And the standard deviation for the individual value is $0.9664\sqrt{1+0.1854} = 1.0522$.

Since $t_{0.025,12} = 2.1788$, to get the 95% confidence interval for *E(Y)* at *x*=5, we compute

$\hat{Y}_j \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{h_j} = 9.3135 \pm 2.1788 \times 0.4161 = 9.3135 \pm 0.9066$.

To get the 95% confidence interval for *Y* at *x*=6, we compute

$\hat{Y}_j \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1+h_j} = 9.3135 \pm 2.1788 \times 1.0522 = 9.3135 \pm 2.2925$.

The above 95% confidence interval for *E(Y)* at *x*=5 is interpreted as *"With 95% confidence the mean annual sales of all stores with the siz of 5 thousand square feet is between $8.4069 million and $10.2201 million."*
The above 95% prediction interval for *Y* at *x*=5 is interpreted as *"With 95% confidence the annual sales of a store with thte sizs of 5 thousand square feet is between $7.0210 milliion and $11.6060 million."*

MSL Homework: 13.57
HC Homework: 13. 61 (Use Excel/PHStat)

When there are more than one independent variable are considered, we have the **multiple regression** model. For example if two independent variables, X and W are considered, the model is

$$Y = \beta_0 + \beta_1 X + \beta_2 W + \varepsilon .$$

Statistical inference of the multiple regression model will be discussed in MgtOp 412: Statistical Methods for Management or ECONS 311: Introductory Econometrics.

Example: In finance, it is of interest to look at the relationship between a stock's average return in percent (Y) and the overall market return in percent (X). From 12 randomly selected stocks we obtain the following data.

Obs.	market return (X)	stock's return (Y)
1	3.7	10.1
2	5.0	12.7
3	1.2	8.6
4	6.1	15.0
5	3.4	9.0
6	3.9	11.7
7	2.0	8.1
8	2.3	10.4
9	6.1	13.2
10	4.2	11.6
11	3.0	9.7
12	2.7	7.3

1. Find the sample correlation coefficient between X and Y.
2. How would you decide if a simple linear regression model is appropriate for the relationship between X and Y?
3. If a simple linear regression model is indeed appropriate for the relationship, find the estimated regression line.
4. Find the predicted average return of a stock with the overall market return of 3%.
5. Is there strong evidence that the average return of a stock is linearly related to the overall market return? Justify your answer.
6. Find a 95% confidence interval for the slope parameter β_1. Note in finance the slope coefficient β_1 is called the stock's *beta* by investment analysts.
7. A *beta* greater than one indicates that the stock is relatively sensitive to changes in the market, while a beta less than one indicates that the stock is relatively insensitive. For the data analyzed, test if the estimated *beta* is significantly greater than one. Use α=0.05.
8. Find an estimate for the variance of the error in the simple linear regression model.
9. Find the 95% confidence interval for the mean of the average returns of stocks with the market return of 8% and interpret the C.I.
10. Find the 95% prediction interval for the average return of a stock with the market return of 8% and interpret the C.I.
11. Find a 90% confidence interval for the mean of the average returns of stocks with the market return of 8%.
12. Find a 90% prediction interval for the average return of a stock with the market return of 8%.