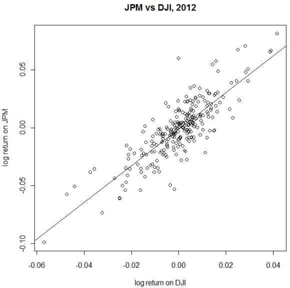
The Simple Linear Regression Model

A Linear Probabilistic Model

Regression Anslysis

- In practice we always observe more than one variables. We need to exploit the relationship between these variables so that we can gain information about one of them through knowing the value of the others.
- This relationship maybe non-deterministic.



Log return
$r_t = log(S_t) - log(S_{t-1})$.

$r_{JPM} = -0.0014 + 1.57 * r_{DJI}$,
$R^2 = 0.689$

Two Variables

- For simplicity, we only consider two variable, x and y
- The simplest relationship is linear relationship $y = \beta_0 + \beta_1 x$
- x is called **independent variable, predictor** or **explanatory variable**.
- y is called **dependent variable** or **response variable**.
- The available data consist of n pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$.
- A picture of this data called a **scatter plot** gives preliminary impressions about the nature of any relationship. This may give us clue to find relationships.

A Linear Probabilistic Model

There are parameters β_0, β_1 and σ^2, such that for any fixed value of the independent variable x, the dependent variable is a random variable related to x through the **model equation**

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

The quantity ϵ in the model equation is a random variable, assumed to be normally distributed with

$$E(\epsilon) = 0, \quad Var(\epsilon) = \sigma^2.$$

A Linear Probabilistic Model

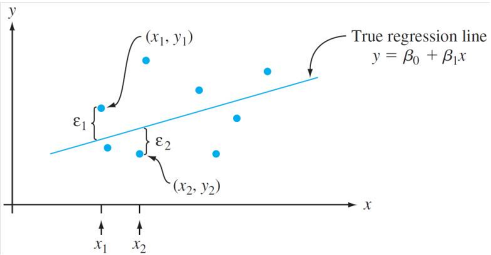
The variable ϵ is usually referred to as the bf random deviation or **random error term** in the model.

Without ϵ, any observed pair (x, y) would correspond to a point falling exactly on the line $y = \beta_0 + \beta_1 x$, called the **true (or population) regression line**.

The inclusion of the random error term allows (x, y) to fall either above the true regression line (when $\epsilon > 0$) or below the line (when $\epsilon < 0$).

A Linear Probabilistic Model

The points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ resulting from n independent observations will then be scattered about the true regression line.



A Linear Probabilistic Model

Once x is fixed, the only randomness on the right-hand side of the model equation is in the random error ϵ, and its mean value and variance are 0 and σ^2, respectively, whatever the value of x.

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x$$

$$Var(Y|x) = Var(\beta_0 + \beta_1 x + \epsilon) = Var(\epsilon) = \sigma^2$$

Interpretation

- The true regression line $y = \beta_0 + \beta_1 x$ is thus the line of mean values; its height above any particular x value is the expected value of Y for that value of x
- The slope β_1 of the true regression line is interpreted as the expected change in Y associated with a 1-unit increase in the value of x.