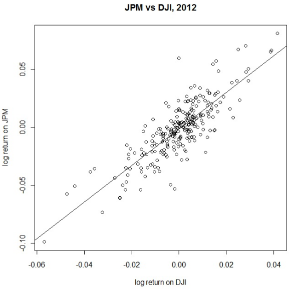
The Simple Linear Regression Model

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

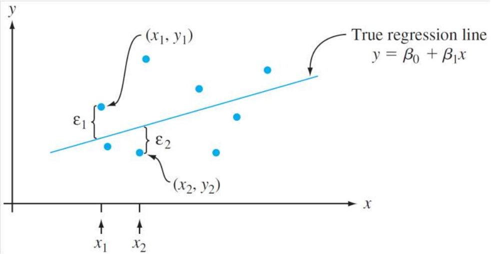
The variable ϵ is usually referred to as the bf random deviation or **random error term** in the model.

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.



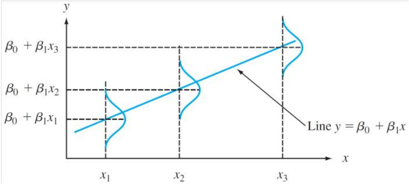
A Linear Probabilistic Model

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.

$$\begin{aligned} E(Y|x) &= E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x \\ Var(Y|x) &= Var(\beta_0 + \beta_1 x + \epsilon) = Var(\epsilon) = \sigma^2 \end{aligned}$$

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.



Estimating Model Parameters

Least Squares

Estimating σ^2

The Coefficient of Determination

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

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

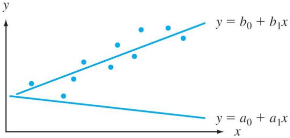
Parameter Estimation

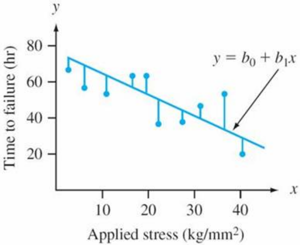
- Sample data consists of n pairs $(x_1, y_1), \cdots, (x_n, y_n)$
- These observations are assumed to have been obtained independently of one another.

  That is, y_i is the observed value of Y_i, where $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and the n deviations $\epsilon_1, \epsilon_2, \epsilon_n$ are independent rv's.
- The values of β_0, β_1, and σ^2 remain to be estimated.

Least Squares

- A good estimate for regression line should be the one that observed points are scattered rather closely about this line.
- A measure of closeness is the vertical distances (deviations) from the observed points to the line.
- The best-fit line is then the one having the smallest possible sum of squared deviations.





Principle of Least Squares (Ordinary Least Squares or OLS)

- The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1 x$ is

$$\text{height of point} - \text{height of line} = y_i - (b_0 + b_1 x_i)$$

- The sum of squared vertical deviations from the points $(x_1, y_1), \cdots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

- The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least squares estimates**, are those values that minimize $f(b_0, b_1)$.
- The estimated regression line or least squares line is then the line whose equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $f(b_0, b_1)$ with respect to both b_0 and b_1, equating them both to zero [analogously to $f'(b) = 0$ in univariate calculus], and solving the equations

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$