Dependence Between Two Samples

- Two independent sample. $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ are independent, e.g. SAT scores of students from two different high schools.
- Two dependent sample. $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ are dependent, e.g. Math scores and Physics scores of students who are taking both these exams, in which there is natural pairing of values.

Example

In a study, six river locations were selected (six experimental objects) and the zinc concentration (mg/L) determined for both surface water and bottom water at each location.

The six pairs of observations are displayed in the accompanying table. Does the data suggest that true average concentration in bottom water exceeds that of surface water?

	Location					
	1	**2**	**3**	**4**	**5**	**6**
Zinc concentration in bottom water (x)	.430	.266	.567	.531	.707	.716
Zinc concentration in surface water (y)	.415	.238	.390	.410	.605	.609
Difference	.015	.028	.177	.121	.102	.107

Example Cont'd

At first glance, the data appears to be little difference between the x and y samples.

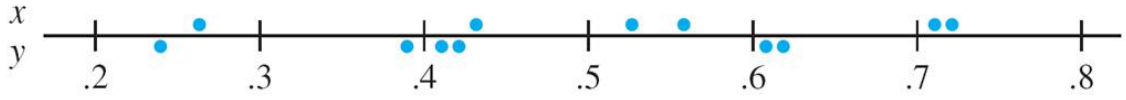


Figure: observations not identified by location

From location to location, there is a great deal of variability in each sample. It looks any differences between the samples can be attributed to location variability.

Example Cont'd

However, when the observations are identified by location, a different view emerges. At each location, bottom concentration exceeds surface concentration.

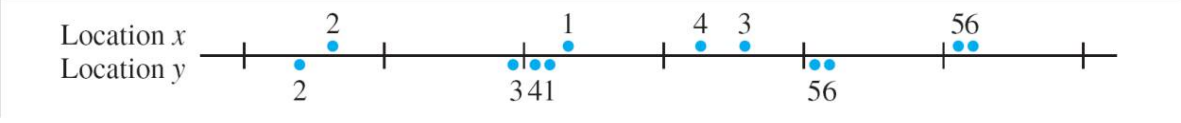


Figure: observations identified by location

This is confirmed by the fact that all $x - y$ differences are positive.

A correct analysis of this data focuses on these differences.

Assumptions

- The data consists of n independently selected pairs $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$.
- $E(X_i) = \mu_1$, $E(Y_i) = \mu_2$.
- Let $D_1 = X_1 - Y_1$, ..., $D_n = X_n - Y_n$ be the differences within pairs.
- Assume D_i's are normally distributed as $N(\mu_D, \sigma_D^2)$

The Paired t Test

- We are again interested in making an inference about the difference $\mu_D = \mu_1 - \mu_2$. An natural estimator will be $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \bar{X} - \bar{Y}$.
- $E(\bar{D}) = E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2 = \mu_D$
- For paired data $\bar{X}$ and $\bar{Y}$ are no longer independent, so $Var(\bar{X} - \bar{Y}) \neq Var(\bar{X}) + Var(\bar{Y})$.
- Since the D_i's constitute a normal random sample (of differences) with mean μ_D, hypotheses about μ_D can be tested using a one-sample t test.
- That is, to test hypotheses about $\mu_1 - \mu_2$ when data is paired, form the differences $D_1, D_2, \cdots, D_n$ and carry out a one-sample t test (based on $n - 1$ df) on these differences.

The Paired t Test

$D = X - Y$ is the difference between observations within a pair. $\mu_D = \mu_1 - \mu_2$. $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ and $s_D = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})$ are sample mean and sample sd of d_i's.

- Null hypothesis: $H_0 : \mu_D = \Delta_0$
- Test statistic value: $t = \frac{\bar{d} - \Delta_0}{s_D/\sqrt{n}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
$H_a : \mu > \mu_0$	$t \geq t_\alpha(n-1)$
$H_a : \mu < \mu_0$	$t \leq -t_\alpha(n-1)$
$H_a : \mu \neq \mu_0$	$t \geq t_{\alpha/2}(n-1)$ or $t \leq -t_{\alpha/2}(n-1)$

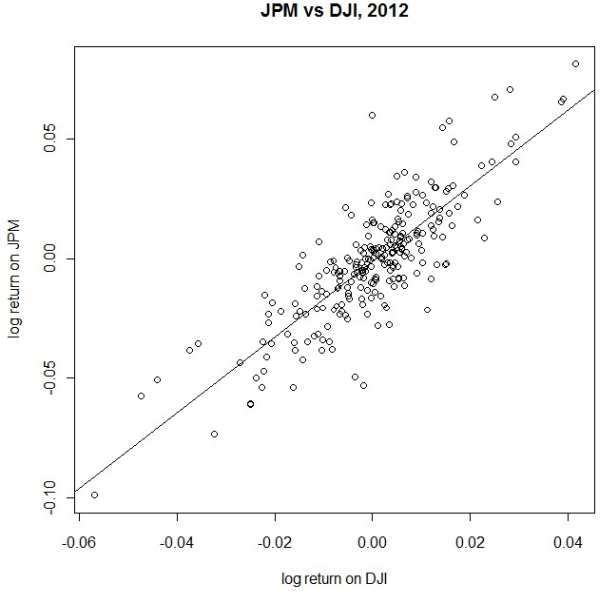
A P-value can be calculated as was done for one sample t test.

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

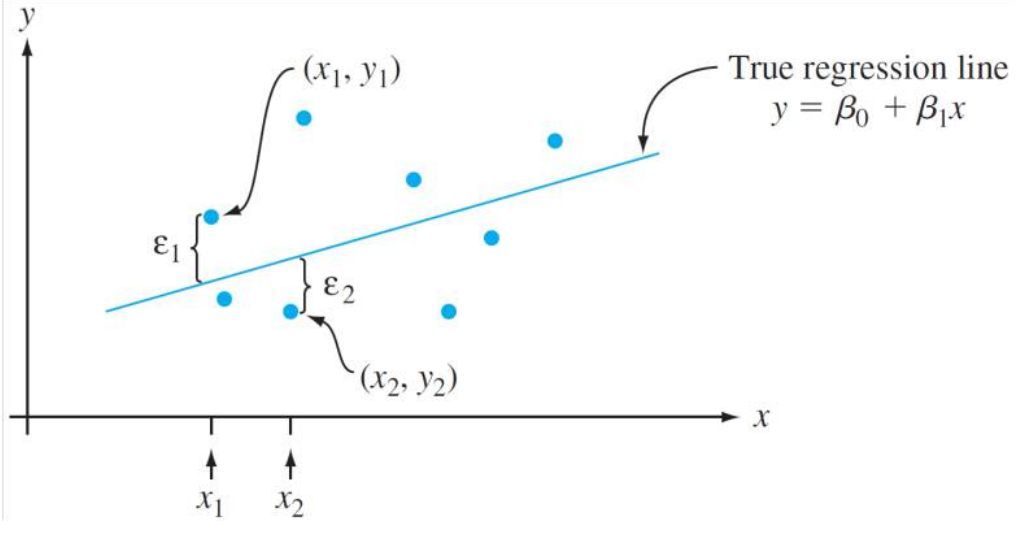
The variable ϵ is usually referred to as the bf random deviation or **random error term** in the model.

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.