**Chapter 13: Simple Linear Regression**

**Assumptions of a Simple Regression Model**

Assumption 1: The random error term $ϵ$ has a mean equal to zero for each $x$.



This means that on average the error terms is equal to 0 for each value of $x$.

Assumption 2: The errors associated with different observations are independent. The error value for $x\_{1}$ should not influence the error value for $x\_{2}$.

Assumption 3: For any given *x*, the distribution of errors is normal.



The mean of normal distribution is 0, as is the mean of errors by Assumption 1. The normal distribution further implies that the concentration of the error terms are towards the center, i.e. near the regression line, with fewer error terms further from the center on either sides.

Assumption 4: The distribution of population errors for each *x* has the same (constant) standard deviation, which is denoted σ*Є .*

This means the spread of observations around the regression line for each value of $x$ is the same. This property is called homoscedasticity. If this assumption is violated then the errors are said to be heteroskedastic.



Assumption 5: The model must be linear in parameters. This means that B cannot be squared or cubed etc.

$\hat{y}=A+B^{2}x$ : such form of regression model is not allowed.

Assumption 6: All the values of x cannot be the same – at least two values of x must be different. Otherwise, the values of $SS\_{xy}$ and $SS\_{xx}$ becomes undefined.

Assumption 7: The values of x must be randomly selected. This means that one value of x must not affect what the next value of x will be.

Assumption 8: The error term cannot be correlated with the independent variable. If this assumption is violated then the regression model suffers from Endogeneity Bias.

Standard Deviation of Random Errors

To calculate the standard deviation of error terms, apply the following formula.

**Coefficient of Determination**

The **coefficient of determination** (denoted by R2) is a key output of [regression](http://stattrek.com/Help/Glossary.aspx?Target=Regression) analysis. It is interpreted as the proportion of the variation in the dependent variable that is predictable from the independent variable, i.e. how well does the independent variable predict the changes in the dependent variable.

* The coefficient of determination is the square of the [correlation](http://stattrek.com/Help/Glossary.aspx?Target=Correlation) (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
* With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.
* An R2 of 0 means that the dependent variable cannot be predicted from the independent variable.
* An R2 of 1 means the dependent variable can be predicted without error from the independent variable.
* An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

To measure the coefficient of determination, we need to know the three components SSE, SST and SSR.

* **Total Sum of Squares (SST)**: this measures the total squared distance of each observation value of y from the mean of y.

In this example, we are modeling the relationship between income (x) and food expenditure (y). Information on income and expenditure on food consumption has been collected for 7 individuals are presented in a scatter diagram. The average food expenditure (from these 7 observations) is $15.43. SST is the sum of the squared difference between each of the 7 points and the mean value.



* **Regression Sum of Squares (SSR)**: this measures the total squared distance from the regression line to the mean of y.
* **Error Sum of Squares (SSE)**: this measures the total squared distance from the observation points to the regression line.

$$SST=SSR+SSE$$



Hence total variation from the mean to Y7 is SST. SSR measures how much of the variation of Y7 from the mean is explained by the regression line. The proportion of SST that is explained by SSR is called the coefficient of determination, $R^{2}$ (or your book uses $r^{2}$).

The ***coefficient of determination***, denoted by *r2*, represents the proportion of SST that is explained by the use of the regression model. The computational formula for *r2* is

* and 0 ≤ *r2* ≤ 1

Inferences about B

Recall that the regression model with sample data is $\hat{y}=a\pm bx$. $b$ has certain special properties.

Let the mean of $b$ be and the standard deviation of $b$ be



Hypothesis Testing for B

Suppose we are investigating the relationship between Income (x) and Food Expenditure (y). The relationship is expected to be positive, i.e. B > 0.

* *H*0: *B* = 0 (The slope is zero)
* *H*1: *B* > 0 (The slope is positive)

To test this hypothesis, we collect information on income and food expenditure of 10 households. From this, we calculate b and the standard deviation of b.

Assume population standard deviation is unknown. Hence the hypothesis test will follow the t-distribution.



Conduct a regular right-tailed hypothesis test using the t-distribution.