Simple Linear Regression Model

Assumptions Underlying the Method of Least Squares

In regression analysis our objective is not only to estimate β_0 and β_1 , but also to draw inferences about the true β_0 and β_1. For example, we would like to know how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to their counterparts in the population or how close $\hat{Y}_i$ is to the true $E(Y \mid X_i)$. We must not only specify the functional form of the **model (3.1)**, but also make certain assumptions about the manner in which Y_i are generated.

To see why this requirement is needed, recall **model (3.1)**. It shows that Y_i depends on both X_i and ε_i. Therefore, unless we are specific about how X_i and ε_i are created or generated, there is no way we can make any statistical inference about the Y_i and also, as we shall see, about β_0 and β_1. Thus, the assumptions made about the X_i variable(s) and the error term are extremely critical to the valid interpretation of the regression estimates.

Assumption 1: The regression model is linear in the parameters.

Assumption 2: Values taken by the independent variable X are considered fixed in repeated samples.

Assumption 3: No significant outliers.

Assumption 4: Homoscedasticity or equal variance of error term ε_i.

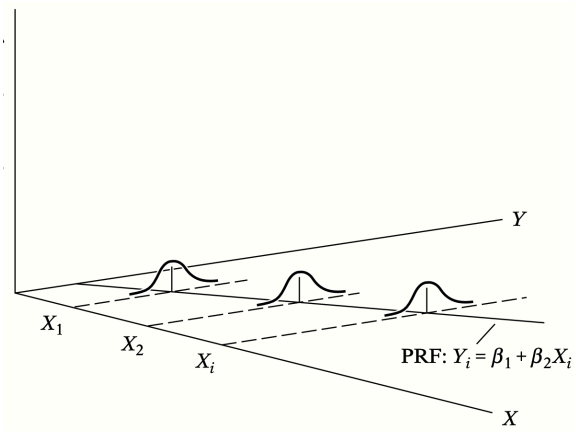

Figure1. Homoscedasticity

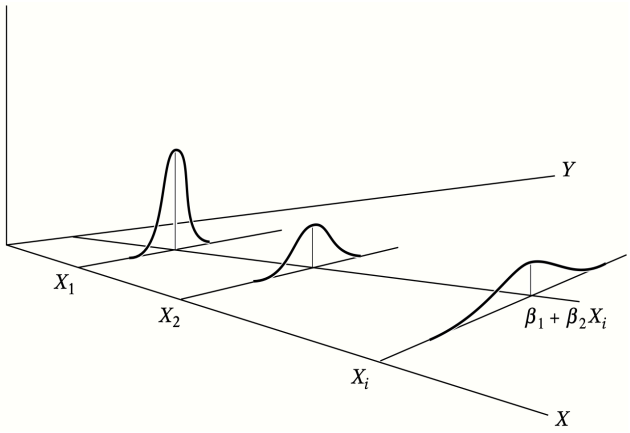

Figure2. Homoscedasticity

Assumption 5: No autocorrelation between the error term ε_i. $\sigma(\varepsilon_i, \varepsilon_j) = 0$ for all $i, j : i \neq j$.

Assumption 6: Zero covariance between ε_i and X_i, or $E(\varepsilon_i X_i) = 0$.

Show that $\sigma(\varepsilon_i, X_i) = 0$.

Assumption 7: The number of observations n must be greater than the number of parameters to be estimated.

Assumption 8: Variability in X values.

Assumption 9: The regression model is correctly specified. No specification bias or error in the model.

Assumption 10: There is no multicollinearity.

Assumption 11: The error terms are normally distributed.

Properties of Least Square Estimators: The Gauss - Markov Theorem

An estimator, say the least square estimator $\hat{\beta}_1$ is said to be a **best linear unbiased estimator (BLUE)** of β_1 if the following hold:

1. It is **linear,** that is, a linear function of a random variable, such as the dependent variable Y in the regression model.
2. It is **unbiased,** that is, its average or expected value, $E(\hat{\beta}_1)$, is equal to the true value, β_1.
3. It has minimum variance in the class of all such linear unbiased estimators; an unbiased estimator with the least variance is known as an **efficient estimator.**

An important theorem, called the **Gauss - Markov Theorem,** states:

> *Under the conditions of regression* ***model (3.1)****, the least square estimators* $\hat{\beta}_0$ *and* $\hat{\beta}_1$*, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.*

Precision or Standard Error of Least Squares Estimates

Measures of precision, variances or standard errors of the estimates, provide a basis for judging the reliability of the estimates. From **(3.8)** and **(3.9)**, it is evident that least-squares estimates are a function of the sample data. But since the data are likely to change from sample to sample, the estimates will change automatically. Therefore, what is needed is some measure of "reliability" or

precision of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. In statistics the precision of an estimate is measured by its standard error (se).

Take Note!!!

The computed regression coefficients, the $\hat{Y}_i$, and the residuals are **all linear functions of the** Y_i. Their variances can be determined using the basic definition of the variance of a linear function.

We need to recognize that $\hat{\beta}_1$ is a linear combination of the observation Y_i:.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n} k_i Y_i$$

Let $\hat{\beta}_1 = \sum_{i=1}^{n} k_i Y_i$ be an arbitrary linear function of the random variables Y_i, where the $k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ are constants.

Observe that k_i are a function of the X_i and therefore fixed quantities since the X_i are fixed. Hence, $\hat{\beta}_1$ is a linear combination of the Y_i where the coefficients are solely a function of the fixed X_i.

The coefficients k_i have a number of interesting properties that will be used.

1. $\sum_{i=1}^{n} k_i = 0$

2. $\sum_{i=1}^{n} k_i x_i = 1$

3. $\sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

The **standard error of the $\hat{\beta}_1$:**

$$\hat{\sigma}(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \qquad \textbf{(1.13)}$$

The variance of $\hat{\beta}_1$ is directly proportional to $\hat{\sigma}^2$ but inversely proportional to $\sum_{i=1}^{n}(X_i - \bar{X})^2$. That is, given $\hat{\sigma}^2$, the larger the variation in the X values, the smaller the variance of $\hat{\beta}_1$ and hence the greater the precision with which β_1 can be estimated.

Exercise:

Derive the standard error of the estimated regression slope coefficient **(1.13)**.

We need to recognize that $\hat{\beta}_0$ is a linear combination of the observation d_i:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = \sum_{i=1}^{n} d_i Y_i$$

Let $\hat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i$ be an arbitrary linear function of the random variables Y_i, where the $d_i = \frac{1}{n} - \bar{X}k_i$.

The coefficients d_i have a number of interesting properties that will be used.

1. $\sum d_i = 1$

2. $\sum_{i=1}^{n} d_i x_i = 0$

3. $\sum_{i=1}^{n} d_i^2 = \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

The **standard error of the $\hat{\beta}_0$:**

$$\hat{\sigma}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \qquad \textbf{(3.14)}$$

Exercise:

Derive the standard error of the estimated regression intercept coefficient **(3.14)**.

Estimator of σ^2

It seems reasonable to assume that the greater the variability of the random error ε (which is measured by its variance σ^2), the greater will be the errors in the estimation of the model parameters β_0 and β_1, and in the error of prediction when $\hat{Y}_i$ is used to predict Y_i for some value of X_i.

In most practical situations, σ^2 will be unknown, and we must use the data to estimate its value. The best estimate of σ^2 is $\hat{\sigma}^2$, which is obtained by dividing the **sum of squares of residuals (SSE)**

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad \textbf{(3.15)}$$

by the number of **degrees of freedom (df)** associated with this quantity. We use 2 df to estimate the y-intercept and slope in the linear regression model, leaving $(n-2)$ df for the error variance estimation.

- **Variance Error of the Estimate,** σ^2

$$\sigma^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n} = \frac{\sum_{i=1}^{n} (Y_i - E(Y_i))^2}{n} \tag{3.16}$$

- **Variance Error of the Estimator,** $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2} \tag{3.17}$$

- **Standard Error of the Estimator or Standard Error of the Regression,** $\hat{\sigma}$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}} \tag{3.18}$$

  (3.18) is simply the standard deviation of the Y_i values about the estimated regression line and is often used as a summary measure of the *"goodness of fit"* of the estimated regression line.

- **Sample Variance of X,** $\hat{\sigma}_X^2$

$$\hat{\sigma}_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}_i)^2}{n-1}$$

- **Sample Variance of Y,** $\hat{\sigma}_Y^2$

$$\hat{\sigma}_Y^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2}{n-1}$$

  Take the square root of the sample variance to get the sample standard deviation of X $(\hat{\sigma}_X)$ and sample standard deviation of Y $(\hat{\sigma}_Y)$.

Exercise:

No.	No. Of Hours Studying (X)	Final Scores (Y)	$\hat{Y}$	e	e^2
1	2	21			
2	4	27			
3	6	29			
4	8	64			
5	10	86			
6	12	92			
Total	**42**	**319**			

A. Obtain the least squares estimates of β_0 and β_1.

B. Write the equation of the regression line.

C. Using the data in the previous Exercises, complete the information:

D. Compute the following:

D.1 $\hat{\sigma}_Y^2$

D.2 $\hat{\sigma}_X^2$

D.3 $\hat{\sigma}^2$

D.4 $\hat{\sigma}$

D.5 $\hat{\sigma}(\hat{\beta}_o)$

D.6 $\hat{\sigma}(\hat{\beta}_1)$

Assignment:

1. Prove that the LSE's $\hat{\beta}_0$ and $\hat{\beta}_1$ of the regression coefficients β_0 and β_1, are BLUE (best Linear Unbiased Estimators).
2. Derive the LSE and MLE for the error variance, σ^2 and show that it is an unbiased estimator.

Prepared by:

KATRINA D. ELIZON
Department of Mathematics and Statistics
College of Science