Applied Mathematics, 2012, 3, 1945-1949
http://dx.doi.org/10.4236/am.2012.312266 Published Online December 2012 (http://www.SciRP.org/journal/am)

Generalized Minimum Perpendicular Distance Square Method of Estimation

Rezaul Karim, Morshed Alam, M. M. H. Chowdhury, Forhad Hossain
Department of Statistics, Jahangirnagar University, Savar, Bangladesh
Email: rezaul@juniv.edu, Morshed@juniv.edu, munirhc@yahoo.com, forhad.ju88@yahoo.com

Received September 22, 2012; revised October 22, 2012; accepted October 30, 2012

ABSTRACT

In case of heteroscedasticity, a Generalized Minimum Perpendicular Distance Square (GMPDS) method has been suggested instead of traditionally used Generalized Least Square (GLS) method to fit a regression line, with an aim to get a better fitted regression line, so that the estimated line will be closest one to the observed points. Mathematical form of the estimator for the parameters has been presented. A logical argument behind the relationship between the slopes of the lines $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and $\hat{X}_i = \hat{\beta}'_0 + \hat{\beta}'_1 Y_i$ has been placed.

Keywords: Heteroscedasticity; Ordinary Least Square Method; Minimum Perpendicular Distance Square Method; Generalized Least Square Method

1. Introduction

Linear regression has a long history in its way of development from the very begging of eighteenth century till today. A lot of literatures are available in this area, these literatures involves the estimation of regression coefficients and constant by Ordinary Least Square (OLS) method *i.e.* by minimizing the sum of square of the vertical distances between the observed points and the assumed regression line, and estimate the regression coefficients traditionally known as OLS estimation procedure.

M. F. Hossain and G. Khalaf, (2009) showed that OLS method does not minimize actual distance from the observed point to the fitted regression line. They have suggested minimum perpendicular distance square (MPDS) Method estimation for simple linear regression in case of homoscedasticity which boils down the traditional OLS method. But regression disturbances whose variances are not constant across observations are heteroscedastic. Heteroscedasticity arises in numerous applications, in both cross-section and time-series data. For example, even after accounting for firm sizes, we expect to observe greater variation in the profits of large firms than in those of small ones. The variance of profits might also depend on product diversification, research and development expenditure, and industry characteristics and therefore might also vary across firms of similar sizes. When analyzing family spending patterns, we observe greater variation in expenditure on certain commodity groups among high-income families than low ones due to the greater discretion allowed by higher incomes [1]. MPDS method is not suitable for this type of heteroscedasticity situation because this method was established only for homoscedasticity cases.

In this paper we have considered minimum perpendicular distance square method in case of heteroscedasticity which we called Generalized Minimum Perpendicular Distance Square (GMPDS) method.

2. Problems of Ordinary Least Square (OLS) and Generalized Least Square (GLS) Method

Suppose the simple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where the response variable Y is related to the explanatory variable X through the regression coefficient β_1, constant intercept β_0 and random disturbance term u. We assume that the disturbance terms u_i follow all assumptions of classical linear regression model.

The estimation procedure of regression coefficient by Ordinary Least Square (OLS) method and Generalized Least Square (GLS) method is actually *minimizing the sum of square of the vertical distances* $(\tilde{u}_i)$ from the observed points to the assumed regression line.

The OLS estimators are:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{SPxy}{SSx}$$

and $\tilde{\beta}_0 = \bar{Y} - \tilde{\beta}_1 \bar{X}$.

The important assumption for applying OLS method is that the variance of each disturbance term u_i, conditional on the chosen values of the explanatory variables, is some constant number (is called homoscedasticity assumption). If the data violet this homoscedasticity assumption that is the variance of each disturbance term u_i conditional on the chosen values of the explanatory variables is random (say σ_i^2) then we can not apply OLS and in this case we apply GLS estimation procedure for estimating parameters [2].

The GLS estimators are:

$$\tilde{\beta}_1^* = \frac{\left(\sum w_i\right)\left(\sum w_i X_i Y_i\right) - \left(\sum w_i X_i\right)\left(\sum w_i Y_i\right)}{\left(\sum w_i\right)\left(\sum w_i X_i^2\right) - \left(\sum w_i X_i\right)^2}$$

and $\tilde{\beta}_0^* = \bar{Y}^* - \tilde{\beta}_1^* \bar{X}^*$

where,

$$w_i = \frac{1}{\sigma_i^2}, \bar{X}^* = \frac{\sum w_i X_i}{\sum w_i}$$

$$\text{and } \bar{Y}^* = \frac{\sum w_i Y_i}{\sum w_i}$$

The problem of OLS and GLS estimation is that, actually they don't minimize real distance from the observed point to the fitted regression line rather they minimize the vertical distance from the observe point to the fitted regression line. For this reason we have the well known theorem is

$$\gamma^2 = \tilde{\beta}_{XY} \tilde{\beta}_{YX} .$$

where $\tilde{\beta}_{XY}$ is the estimated regression coefficient of X on Y and $\tilde{\beta}_{YX}$ is the estimated regression coefficient of Y on X. If OLS and GLS minimize real distance (error) then $\tilde{\beta}_{XY}\tilde{\beta}_{YX}$ should be unity that is $\tilde{\beta}_{XY}\tilde{\beta}_{YX} = 1$. But in OLS and GLS methods, it only occurs if data are perfectly correlated, that is $r = \pm 1$. In real life problem this type of perfect correlation occurs in rare case.

The Minimum Perpendicular Distance Square Method suggested by Hossain and Khalaf (2009) produced the estimator which gives $\hat{\beta}_{XY}\hat{\beta}_{YX} = 1$ for all cases and it indicates that the errors are really minimized and gives more accurate result than that of OLS [3].

Concept of Minimum Perpendicular Distance Square (MPDS) Estimation

The real distance of the assumed regression line $Y = \beta_0 + \beta_1 X$ from the points $\left(X_i, Y_i\right); i = 1, 2, \cdots, n$ are not the vertical distances or height of the point minus height of regression line *i.e.* $Y_i - \left(\beta_0 + \beta_1 X_i\right)$.

In fact the actual distances from the line $Y = \beta_0 + \beta_1 X$ to the points $\left(X_i, Y_i\right); i = 1, 2, \cdots, n$ are the perpendicular distances $\hat{u}_i$ *'s* (as indicated in **Figure 1**). These perpendicular distances would also be positive and negative according to $\left(X_i, Y_i\right)$ is above the line $\left(\hat{u}_i > 0\right)$ or below the line $\left(\hat{u}_i < 0\right)$. Also assuming that $\hat{u}_i \sim N\left(0, \left(\sigma'\right)^2\right)$. Hence estimating β_1 and β_0 by minimizing sum of the squares of these perpendicular distances will produce the closest fitted regression line from the points $\left(X_i, Y_i\right); i = 1, 2, \cdots, n$ which may be used for more accurate prediction purposes.

3. The Method of Generalized Minimum Perpendicular Distance Squares Method (GMPDSM)

Let us consider two-variable linear regression function is

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

which for ease of algebraic simplification we write as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_i + u_i \tag{1}$$

where $X_{0i} = 1$ for each i and the response variable Y is related to the explanatory variable X through the regression coefficient β_1, constant intercept β_0 and random disturbance term u. We know that one of the important assumptions of the classical linear regression model is that the variance of each disturbance term u_i, conditional on the chosen values of the explanatory variables is some constant number equal to σ^2. This is the assumption of homoscedasticity. Symbolically,

$$Var\left(u_i\right) = E\left(u_i^2\right) = \sigma^2 ; i = 1, 2, \cdots, n$$

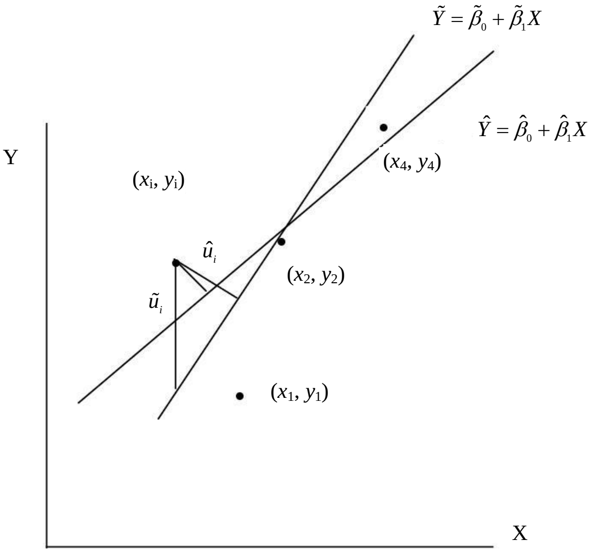
$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 X$
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
Y
(x_4, y_4)
(x_i, y_i)
$\hat{u}_i$
(x_2, y_2)
$\tilde{u}_i$
(x_1, y_1)
X

Figure 1. Regression lines obtained from OLS & MPDS method.

Now if the conditional variance of $Y_i\left(\text{or } u_i\right)$ are not same for each of the u_i. *i.e.*, heteroscedasticity. Symbolically,

$$Var\left(u_i\right)=E\left(u_i^2\right)=\sigma_i^2\,;i=1,2,\cdots,n$$

and suppose the heteroscedastic variance σ_i^2 are *known*. Then dividing (1) by σ_i^2 both sides, we get

$$\frac{Y_i}{\sigma_i}=\beta_0\left(\frac{X_{0i}}{\sigma_i}\right)+\beta_1\left(\frac{X_i}{\sigma_i}\right)+\left(\frac{u_i}{\sigma_i}\right) \qquad (2)$$

which for ease of exposition we write

$$Y_i^*=\beta_0^*X_{0i}^*+\beta_1^*X_i^*+u_i^* \qquad (3)$$

where the transformed variables are the original variables divided by (the known) σ_i. We use the notation β_0^* and β_1^*, the parameters of the transformed model, to distinguish them from the usual MPDS parameters β_0 and β_1. Now we see

$$Var\left(u_i^*\right)=E\left(u_i^*\right)^2=E\left(\frac{u_i}{\sigma_i}\right)^2=\frac{1}{\sigma_i^2}E\left(u_i^2\right)$$

since σ_i^2 is known $=1$ since $E\left(u_i^2\right)=\sigma_i^2$

which is a constant. That is, the variance of the transformed disturbance term u_i^* is now homoscedastic.

This procedure of transforming the original variables is done in such a way that the transformed variables satisfy the assumptions of the classical model. Now applying MPDS method to this transformed model to estimate parameter we call Generalized Minimum Perpendicular Distance Squares Method (GMPDSM). *In short, GMPDS is MPDS on the transformed variables that satisfy the classical regression assumptions*. The estimators thus obtained are knows as GMPDSM estimators.

3.1. Perpendicular Distance from the Points to the Line $\hat{Y}_i^*=\hat{\beta}_0^*+\hat{\beta}_1^*X_i^*$

Let us consider two-variable linear regression function

$$Y_i=\beta_0+\beta_1X_i+u_i$$

Dividing both sides by σ_i we have

$$\frac{Y_i}{\sigma_i}=\beta_0^*\left(\frac{X_{0i}}{\sigma_i}\right)+\beta_1^*\left(\frac{X_i}{\sigma_i}\right)+\left(\frac{u_i}{\sigma_i}\right) \qquad (4)$$

or

$$Y_i^*=\beta_1^*X_{0i}^*+\beta_2^*X_i^*+u_i^*$$

For estimating β_0^* and β_0^* we need to determine the perpendicular distance from the observed point $\left(X_i^*,Y_i^*\right)$ to the line $\hat{Y}_i^*=\hat{\beta}_1^*X_{0i}^*+\hat{\beta}_2^*X_i^*$. The perpendicular distance $\hat{u}_i^*$ from the points $\left(X_i^*,Y_i^*\right)$ to the fitted line $\hat{Y}^*=\hat{\beta}_0^*X_{0i}^*+\hat{\beta}_1^*X^*$ [4,5] is

$$u_i^*=\sqrt{\left(X_i^*-\bar{X}^*\right)^2+\left(Y_i^*-\bar{Y}^*\right)^2}$$
$$=\frac{\left(Y_i^*-\beta_0^*X_{0i}^*-\beta_1^*X_i^*\right)}{\pm\sqrt{\beta_1^{*2}+1}}$$

3.2. Parameter Estimation Based on GMPDS Method

To obtain the GMPDS estimators, we minimize sum of square of perpendicular distances $\hat{u}_i^*$ from the points $\left(X_i^*,Y_i^*\right);i=1,2,\cdots,n$ to the fitted line $\hat{Y}=\hat{\beta}_0^*+\hat{\beta}_1^*X^*$ following steps are taken.

$$\sum_{i=1}^{n}\left(u_i^*\right)^2=\sum_{i=1}^{n}\frac{\left(Y_i^*-\beta_0^*X_{0i}^*-\beta_1^*X_i^*\right)^2}{\beta_1^{*2}+1}$$

that is,

$$\sum_{i=1}^{n}w_iu_i^2=\sum_{i=1}^{n}\frac{w_i\left(Y_i-\beta_0^*-\beta_1^*X_i\right)^2}{\beta_1^{*2}+1} \qquad (5)$$

where weights

$$w_i=\frac{1}{\sigma_i^2}$$

that is, the weights are inversely proportional to the variance of u_i or Y_i conditional on the given X_i, *i.e.*, $\text{var}\left(u_i\left|X_i\right.\right)=\text{var}\left(Y_i\left|X_i\right.\right)=\sigma_i^2$.

Differentiating (5) with respect to β_1^*, then putting equal to zero and setting for $\beta_1^*=\hat{\beta}_1^*$ we get the normal equation

$$\begin{aligned}&-\sum w_iX_iY_i+\hat{\beta}_1^*\sum w_iX_i^2+\hat{\beta}_0^*\sum w_iX_i\\&+\hat{\beta}_1^{*2}\sum w_iX_iY_i-\hat{\beta}_0^*\hat{\beta}_1^{*2}\sum w_iX_i\\&-\hat{\beta}_1^{*2}\sum w_iY_i^2-\hat{\beta}_0^{*2}\hat{\beta}_1^*\sum w_i\\&+2\hat{\beta}_0^*\hat{\beta}_1^*\sum w_iY_i=0\end{aligned} \qquad (6)$$

Again differentiating Equation (5) with respect to β_0^* and equating zero with $\beta_0^*=\hat{\beta}_0^*$, we get

$$\left.\frac{\mathrm{d}\sum w_iu_i^2}{\mathrm{d}\beta_0^*}\right|_{\beta_0^*=\hat{\beta}_0^*}=\frac{\sum w_i\left(Y_i-\hat{\beta}_0^*-\hat{\beta}_1^*X_i\right)(-2)}{\left(\hat{\beta}_1^{*2}+1\right)^2}=0$$

$$\Rightarrow\sum w_iY_i+\hat{\beta}_1^*\sum w_iX_i-\hat{\beta}_0^*\sum w_i=0 \qquad (7)$$

$$\Rightarrow\hat{\beta}_0^*=\frac{\sum w_iY_i}{\sum w_i}-\hat{\beta}_1^*\frac{\sum w_iX_i}{\sum w_i}$$

Using Equation (7) in Equation (6) we get

$$\Rightarrow -\sum w_i X_i Y_i + \hat{\beta}_1^* \sum w_i X_i^2$$
$$+\left(\frac{\sum w_i Y_i}{\sum w_i} - \hat{\beta}_1^* \frac{\sum w_i X_i}{\sum w_i}\right)\sum w_i X_i + \hat{\beta}_1^{*2} \sum w_i X_i Y_i$$
$$-\hat{\beta}_1^{*2}\left(\frac{\sum w_i Y_i}{\sum w_i} - \hat{\beta}_1^* \frac{\sum w_i X_i}{\sum w_i}\right)\sum w_i X_i - \hat{\beta}_1^{*2} \sum w_i Y_i^2$$
$$-\hat{\beta}_1^*\left(\frac{\sum w_i Y_i}{\sum w_i} - \hat{\beta}_1^* \frac{\sum w_i X_i}{\sum w_i}\right)^2 \sum w_i$$
$$+2\hat{\beta}_1^*\left(\frac{\sum w_i Y_i}{\sum w_i} - \hat{\beta}_1^* \frac{\sum w_i X_i}{\sum w_i}\right)\sum w_i Y_i = 0$$
$$\text{or, } \beta_1^{*2} SPwxy + \beta_1^* \left(SSwx - SSwy\right) - SPwxy = 0$$

where

$$SSwy = \left\{\sum w_i Y_i^2 - \frac{\left(\sum w_i Y_i\right)^2}{\sum w_i}\right\}$$
$$SSwx = \left\{\sum w_i X_i^2 - \frac{\left(\sum w_i X_i\right)^2}{\sum w_i}\right\}$$
$$SPwxy = \left\{\sum w_i X_i Y_i - \frac{\left(\sum w_i X_i\right)\left(\sum w_i Y_i\right)}{\sum w_i}\right\}$$

So the solution of the above equation is:

$$\hat{\beta}_1^* = \frac{-\left(SSwx - SSwy\right) \pm \sqrt{\left(SSwx - SSwy\right)^2 + 4\left(SPwxy\right)^2}}{2SPwxy}$$

Hence

$$\hat{\beta}_{1(1)}^* = \frac{-\left(SSwx - SSwy\right) + \sqrt{\left(SSwx - SSwy\right)^2 + 4\left(SPwxy\right)^2}}{2SPwxy}$$
$$\hat{\beta}_{1(2)}^* = \frac{-\left(SSwx - SSwy\right) - \sqrt{\left(SSwx - SSwy\right)^2 + 4\left(SPwxy\right)^2}}{2SPwxy}$$

Using this result in Equation (7) we can estimate $\hat{\beta}_0^*$. And hence

$$\therefore \hat{\beta}_0^* = \frac{\sum w_i Y_i}{\sum w_i} - \hat{\beta}_1^* \frac{\sum w_i X_i}{\sum w_i} \quad (8)$$

In this method we get two regression coefficients, it could be proved that the "+" solution *i.e.* $\hat{\beta}_{1(1)}^*$ gives minimum of (5) and hence we suggest the reader to use $\hat{\beta}_{1(1)}^*$ as the regression coefficient and accordingly the regression constant β_0^* could be estimated by using $\hat{\beta}_{1(1)}^*$ in Equation (8) to fit the regression line Y^* on X^* *i.e.* $Y^* = \hat{\beta}_0^* X_0^* + \hat{\beta}_1^* X^*$.

3.3. Estimation of Regression Coefficient by Using GMPDS for the Model $\hat{X}^* = \hat{\beta}_0'^* + \hat{\beta}_1'^* Y^*$

To estimate regression coefficient $\beta_1'^*$ and regression constant $\beta_0'^*$ by minimizing sum of squares of the error term $u_i'^*$'s (assumed) the perpendicular distances from the fitted line $\hat{X}^* = \hat{\beta}_0'^* + \hat{\beta}_1'^* Y^*$ to the points $\left(X_i^*, Y_i^*\right), i = 1, 2, \cdots, n$; we do the similar steps as we do in Section 3.7.

$$\sum_{i=1}^{n}\left(u_i'\right)^2 = \sum_{i=1}^{n}\frac{\left[X_i - \beta_0'^* - \beta_1'^* Y_i\right]^2}{\beta_1'^{*2} + 1}$$

That is,

$$\sum_{i=1}^{n}\left(\frac{u_i'}{\sigma_i}\right)^2 = \sum_{i=1}^{n}\frac{\left[\left(\frac{X_i}{\sigma_i}\right) - \beta_0'^*\left(\frac{1}{\sigma_i}\right) - \beta_1'^*\left(\frac{Y_i}{\sigma_i}\right)\right]^2}{\beta_1'^{*2} + 1} \quad (9)$$
$$\text{or } \sum_{i=1}^{n} w_i u_i'^2 = \sum_{i=1}^{n}\frac{w_i\left(X_i - \beta_0'^* - \beta_1'^* Y_i\right)^2}{\beta_1'^{*2} + 1}$$

Differentiating both sides with respect to $\beta_0'^*$ and $\beta_1'^*$ and putting equal to zero and setting for $\beta_0'^*$ and $\beta_1'^*$, we get the following solutions:

$$\hat{\beta}_1'^* = \frac{-\left(SSwy - SSwx\right) \pm \sqrt{\left(SSwy - SSwx\right)^2 + 4\left(SPwxy\right)^2}}{2SPwxy}$$

Hence

$$\hat{\beta}_{1(1)}'^* = \frac{-\left(SSwy - SSwx\right) + \sqrt{\left(SSwy - SSwx\right)^2 + 4\left(SPwxy\right)^2}}{2SPwxy}$$
$$\hat{\beta}_{1(2)}'^* = \frac{-\left(SSwy - SSwx\right) - \sqrt{\left(SSwy - SSwx\right)^2 + 4\left(SPwxy\right)^2}}{2SPwxy}$$
$$\text{and } \hat{\beta}_0'^* = \frac{\sum w_i X_i}{\sum w_i} - \hat{\beta}_1'^* \frac{\sum w_i Y_i}{\sum w_i}$$

Here we also get two regression coefficients and for the same region as we have mentioned in Section 3.2, we will suggest the reader to use $\hat{\beta}_{1(1)}^{*}{}'$ as regression coefficient and accordingly the estimation of $\hat{\beta}_0^{*\prime}$ may be obtained to fit the regression line X^* on Y^*.

3.4. Relationship between Regression Coefficients

If we consider the GMPDS method to estimate regression coefficients β_1^* and $\beta_1'^*$ as we have indicated in Sections 3.2 and 3.3, by minimizing the error term $\hat{u}_i^*$

and $\hat{u}_i'^*$ respectively (the perpendicular distances from these lines to the observed points), we get

$$\hat{\beta}_{1(1)}^{*} = \frac{-\left(SSwx - SSwy\right) + \sqrt{\left(SSwx - SSwy\right)^2 + 4\left(SPwxy\right)^2}}{2SPwxy}$$

for the line $\hat{Y}^* = \hat{\beta}_0^* X_0^* + \hat{\beta}_1^* X^*$ and

$$\hat{\beta}_{1(1)}'^{*} = \frac{-\left(SSwy - SSwx\right) + \sqrt{\left(SSwy - SSwx\right)^2 + 4\left(SPwxy\right)^2}}{2SPwxy}$$

for the line $\hat{X}^* = \hat{\beta}_0'^* X_0^* + \hat{\beta}_1'^* Y^*$ we see that $\hat{\beta}_{1(1)}^*$ is proportional to $\hat{\beta}_{1(1)}'^*$ *i.e.*

$$\hat{\beta}_{1(1)}^* = \frac{1}{\hat{\beta}_{1(1)}'^*} \text{ or } \hat{\beta}_{1(1)}^* \cdot \hat{\beta}_{1(1)}'^* = 1,$$

which indicate that during estimating regression coefficient by using GMPDS method in case of heteroscedasticity, the error term is minimized. This is a new angle to advocate the advantage our suggested method (GMPDSM) to estimate regression coefficients in case of heteroscedasticity.

4. Concluding Remarks

The method of MPDS estimation actually minimize real distances from the observed points to the fitted regression line but OLS and GLS method fail to do that by using vertical distance from the observe points to the fitted regression line. But one of the crucial assumptions of MPDS method and also for traditional OLS method is that the variance of each disturbance terms remains some constant number $\left(\sigma^2\right)$. So we can not apply MPDS method when this assumption is violated. That is, in presence of heteroscedasticity OLS and MPDS is not suitable. In this paper our main focus is on minimum perpendicular deviations in case of heteroscedasticity, and we have shown in mathematically that GMPDS method gives an estimator that the error term is really minimized. Hence we propose GMPDS method in case of heteroscedasticity.

REFERENCES

[1] W. H. Greene, “Econometric Analysis,” 5th Edition, Pearson Education, Singapore, 2003,

[2] D. Gujarati, “Basic Econometrics,” 4th Edition, McGraw-Hill, New York, 2003.

[3] M. F. Hossain and G. Khalaf, “Minimum Perpendicular Distance Square Method Estimation,” *Journal of Applied Statistical Science*, Vol. 17, No. 2, 2009, pp. 153-180.

[4] A. Mizrahi and M. Sullivan, “Calculus and Analytic Geometry,” Wadsworth Publishing Company, Beverly, 1986.

[5] M. R. Spiegel and John Lin, “Mathematical Handbook of Formulas and Tables,” 2nd Edition, Mcgraw-Hill, New York, 1999.