Finding the Best Fit for Solar Radiation by statistical comparison between Linear Regression (LS) and Least Trimmed Squares (LTS) Regression Method

Y.M. Irwan[1], I. Mohamed[2], A. G. Hussin[2], I.Safwati[2], N.Gomesh[1] and M.Irwanto[1]
[1]Centre of Excellence for Renewable Energy, School of Electrical Systems Engineering, University Malaysia Perlis, 01000, Kangar, Perlis, Malaysia
[2]Institut of Mathematical Science, Faculty of Science, University of Malaya, 50603 Kuala Lumpur

***Abstract*-- The sources of green energy mainly in the area of solar technology have received much emphasis on the global market in terms of power generation. In Malaysia, the research on solar energy has taken a huge leap into power conservation. This paper presents an investigation on solar radiation accuracy in the northern part of Malaysia, the state of Perlis, by using the meteorological data obtain from several weather stations here. The analysis is done by using mathematical and statistical software to compare the air temperature and solar radiation for one complete year. Previous work has done on solar radiation estimation is between Hargreaves and linear regression. Thus paper considers the method of Least Trimmed Squares (LTS) regression in estimating solar radiation. The daily and monthly average data of solar radiation per minute is analyzed by using in Linear Regression (LS) and Least trimmed Squares (LTS) regression method for data comparison reason. Result shows that, by using the LTS method the coefficient of determination, R^2 value is higher compare to the LS method by 7%, other result such as the mean, median and total for the $\hat{y}_{LTS}$ produces similar data as the raw data from weather station. This shows that the LTS method can be used to estimate the solar radiation compare Linear Regression (LS) method. The reason is the LTS method is not easily influence by the outliers of data set. So it can be considered as the best fit for solar radiation model compare to the LS regression method.**

***Index Term*— Photovoltaic, Solar radiation, Air Temperature, Linear regression model (LS), least Trimmed Squares (LTS) robust**

I. INTRODUCTION

AN INCREASE in the use of conventional energy prices and environmental effects, such as air pollution, depletion of the ozone layer and greenhouse effects has made the use of solar energy increased [1]. The availability of more comprehensive solar radiation data is invaluable for the design and evaluation of solar based conversion systems. In many places of the world, particularly the developing countries, the basic solar radiation data for the surfaces of interests are not readily obtainable [2].

Solar radiation is the result of fusion of atoms inside the sun. Part of the energy from the fusion process heats the chromosphere, the outer layer of the sun that is much cooler than the interior of the sun, and the radiation from the chromosphere becomes the solar radiation incident on the earth [3]. Wind energy is produced by continuously blowing wind and can be captured using wind turbines that convert kinetic energy from wind into mechanical energy and then into electrical energy [4].

Fig. 1 shows the solar radiation enters the earth's atmosphere; a part of the incident energy is removed by scattering or absorption by air molecules, clouds and particulate matter usually referred to as aerosols. The radiation that is not reflected or scattered and reaches the surface directly in line from the PV module is called beam radiation. The scattered radiation which reaches the ground is called diffuse radiation. Some of the radiation may reach a receiver after reflection from the ground, and is called the albedo. Albedo is the percentage of incoming radiation reflected off a surface. An albedo of 1 means that 100% of incoming radiation is reflected that mean no radiation is absorbed, meanwhile an albedo of 0 means that 0% of incoming radiation is reflected which is all radiation is absorbed [5].

The total solar radiation on a horizontal surface of PV module consisting three components is called global irradiance. When the skies are clear and the sun is directly in line from the PV module, the global irradiance is about 1000 W/m2 [6]. Although the global irradiance on the surface of the earth can be as high as 1000 W/m2, the available radiation is usually considerably lower than this maximum value due to the rotation on the earth and climate condition (cloud cover), as well as by the general composition of the atmosphere. For this reason, the solar radiation data is the most important component to estimate output of photovoltaic systems [3, 7, & 8]. Solar radiation is greater than 3 kWh/m2 indicates that the sky is clear, its intensity very high and very good for PV application [9].

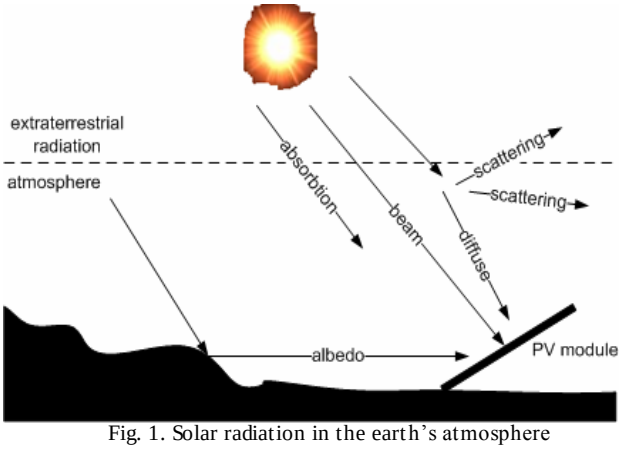

Fig. 1. Solar radiation in the earth's atmosphere

Previous work has been done using various statistical methods such as Hargreaves [10], linear regression model [10] and robust regression model [11]. Since no work has been done on comparing robust least trimmed squares (LTS) regression model, this paper presented the comparison between linear regression and robust LTS regression by using data from solar radiation in Malaysia, the state of Perlis.

II. Description of the Data

Solar-radiation incident on a horizontal surface and sunshine duration are measured by recording stations [12]. Based on Malaysia Meteorological Department [13], Malaysia naturally has abundant sunshine and thus solar radiation. However, it is extremely rare to have a full day with completely clear sky even in periods of severe drought. The cloud cover cuts off a substantial amount of sunshine and thus solar radiation. On the average, Malaysia receives about 6 hours of sunshine per day. Based on Meteorological Station in Chuping, Perlis (60 29' N , 1000 16' E) as shown in Fig 2 has about 795 square kilometers land area, 0.24% of the total land area of Malaysia, with a population about 204450 people [13]. Perlis's climate is tropical monsoon. Its temperature is relatively uniform within the range of 21°C to 32°C throughout the year. During the months of January to April, the weather is generally dry and warm. Humidity is consistently high on the lowlands ranging 82% to 86% per annum. The average rainfall per year is 2,032 mm to 2,540 mm and the wettest months are from May December. In this research, the data are presented in daily averaged maximum and minimum temperature, and daily averaged solar radiation.

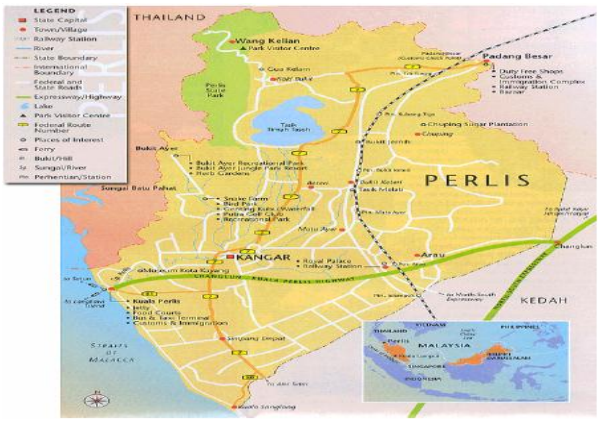

Fig. 2. Map of Perlis has latitude 60° 29' N

III. Methods

1) Simple Linear Regression Model

Regression analysis is a statistical technique for investigating and modelling the relationship between variables [14]. In fact, the regression analysis is the most widely used statistical technique. The simple linear regression model used is a model with a single independent variable x that has a relationship with a response variable y that is a straight line. This simple linear regression model is given by

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where; the intercept β_0 and the slope β_1 are unknown constant and ε is a random error. The errors are assumed to have mean zero and unknown variance σ^2. The parameters β_0 and β_1 are unknown and must be estimated using sample data. The simple linear regression equation is also called the *least squares* (LS) regression equation. It tells the criterion used to select the best fitting line, namely the sum of the *squares* of the residuals should be *least*. That is, the least squares regression equation is the line for which the sum of squared residuals $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is a minimum.

Suppose that for any observation, $(x_1, y_1),(x_2, y_2),...,(x_n, y_n)$ be a pair of random variables. To predict y, the parameters β_0 and β_1 must be estimated, so that the sum of the square of the differences between the observations y_i and the straight line is minimum. The interpolating straight line as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i=1, 2,..., n \qquad (1)$$

and the coefficients that minimise the square of the distance between the line end the points are given by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (2)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

where;

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

are the averages of y_i and x_i, respectively. Therefore, β_0 and β_1 are the least squares estimators of the intercept and slope. The residuals ε are the differences between the

observed and the predicted values $y_i - \hat{y}_i$, i=1, 2, ..., n. The fitted simple linear regression model is given by

$$\hat{y}_{LS} = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (3)$$

The correlation coefficient r evaluates the goodness of the fitting of data considered and the standard error measures, s is calculated. The correlation coefficient value can vary in the range -1 and +1, for the strong correlation between the two variables x and y. If the value is zero there is not any linear correlation between the two variables. The calculation of r and s are respectively as follow

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (4)$$

$$s = \frac{\sum_{i=1}^{n} y_i^{\,2} - B_0\sum_{i=1}^{n} y_i - B_1\sum_{i=1}^{n} x_i y_i}{n} \quad (5)$$

The coefficient of determination, R-Squared $\left(R^2\right)$ is the statistic that will give information about the goodness of fit of the model. R^2 for LS (R_{LS}^2) is given by

$$R_{LS}^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_{i,LS} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,LS}\right)^2} \quad (6)$$

where $0 \le R_{LS}^2 \le 1$. R_{LS}^2 is often called the proportion of variation explained by the regressor x. Values of R_{LS}^2 that are closed to 1 imply that most of the variability in y is explained by the regression model [14].

2) Robust LTS Regression Model

Robust regression is a form of regression analysis devised to overcome some limitations of traditional regression methods. Robust fitting is commonly used when the data contain outliers. The subject occurrence of outliers will be affecting the estimated coefficients, fitted values, residuals and covariance matrix of linear regression models [15, 16]. In the presence of outliers, least squares estimation is inefficient and can be affected by inaccuracy. This is due to the shifting of the least squares estimates towards the outliers and to the corresponding altered increase of the estimates variance. Therefore, the robust approach to statistical modelling and data analysis is produce reliable parameter estimates and associate tests [17]. This method is a best way for identification of parameters solar radiation which do not breakdown easily and are not much influenced by outliers [17].

The parameters identification is performed within SPLUS environment. In particular the embedded robust fit function is used to obtain the regression straight line coefficients estimates. The *ltsreg* function is used in SPLUS. The results are less sensitive to outliers in the data as compared with ordinary least square simple linear regression [5, 16]. One of the robust methods is *least trimmed squares* (LTS) regression. This method is proposed by Rousseeuw. P. J. "Least Median of Squares Regression", [17]. It is a highly robust method for fitting a linear regression model. LTS regression minimizes the sum of the trimmed squared residuals. The estimation of β is obtained from

$$\sum_{i=1}^{h} \hat{\varepsilon}_i^{\,2} \text{ is minimized,}$$

where $\hat{\varepsilon}_1 \le \hat{\varepsilon}_2 \le \hat{\varepsilon}_3 \le \ldots \le \hat{\varepsilon}_n$ are the ordered squared residuals, from smallest to largest. LTS is calculated by minimizing the h ordered squares residuals, where h may depend on a trimming proportion of α, suggested choosing $h = [n(1-\alpha)]+1$. Thus, LTS is equivalent to ordering the residuals from a least squares fit, trimming the observations that correspond to the largest residuals, and then computing a least squares regression model for the remaining observations. The largest squared residuals are excluded from the summation in this method, which allows those outlier data points to be excluded completely. The coefficient of determination, R-Squared for LTS $\left(R_{LTS}^2\right)$ is given by;

$$R_{LTS}^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_{i,LTS} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,LTS}\right)^2} \quad (7)$$

IV. Result and Discussion

1) Data Analysis

Based on the average between maximum and minimum air temperature in Perlis, the solar radiation for the year of 2006 can be estimated using linear regression and least trimmed squares robust regression model. The daily averaged maximum and minimum air temperature and daily averaged solar radiation throughout the year of 2006 in Perlis are shown in Fig 3 and 4, respectively.

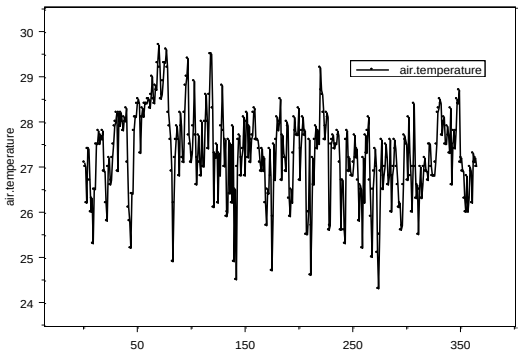

Fig. 3. The graph of average air temperature

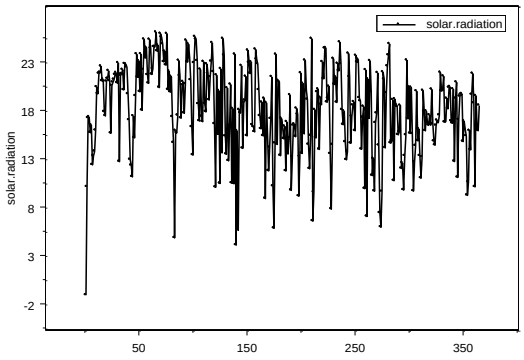

Fig. 4. The graph of solar radiation

2) Parameter Estimation of Least Squares (LS) Method

The analysis part is to find a relation between average solar radiation (*y*) and average air temperature (*x*). The scatter plot in Fig. 5 shows that a strong relationship between these two variables, suggests the possibility to obtain such data by linear regressions.

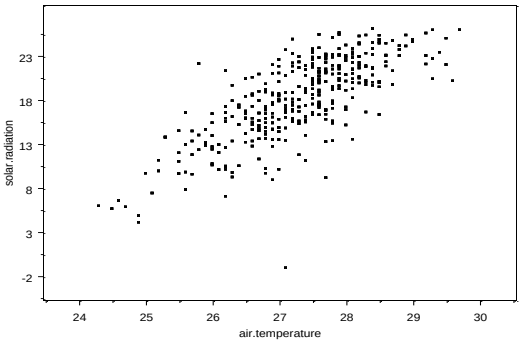

Fig. 5. The scatter plot of solar radiation versus air temperature

A simple linear regression model is assumed, and the estimation of parameters in the regression model is calculated using LS method as given in eq. (2). The value of $\hat{\beta}_0$ and $\hat{\beta}_1$ is -80.4560 and 3.6100, respectively. The least squares fit to the solar radiation data is

$$\hat{y}_{LS} = -80.4560 + 3.6100x$$

where $\hat{y}_{LS}$ is the estimated value of solar radiation

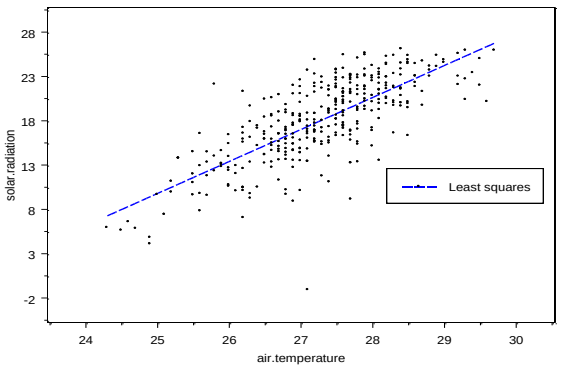

Fig. 6. The scatter plot with the fitted line of air temperature versus solar radiation

The linear correlation coefficient value in this case is 0.7473. It is shows that a strong linear relationship between solar radiation and air temperature. It supported by scatter plot in Fig. 6. This is because the solar radiation and air temperature is directly proportional. The correlation coefficient of determination for LS, $R^2_{LS} = 0.5585$; about 56 percent of the variability in temperature is accounted for by the straight-line fit to solar radiation.

3) Parameter Estimation of Robust LTS Method

The estimation of parameters for robust LTS method is calculated by equation (2). It was found that the $\hat{\beta}_0$ and $\hat{\beta}_1$ are -87.1081 and 3.8594, respectively. Therefore, the least squares regression model using robust LTS is fit to the solar radiation data is

$$\hat{y}_{LTS} = -87.1081 + 3.8594x$$

where $\hat{y}_{LTS}$ is the estimated value of solar radiation corresponding to the air temperature, *x*. The scatter plot with the fitted LTS line of air temperature versus solar radiation as shown in Fig. 7.

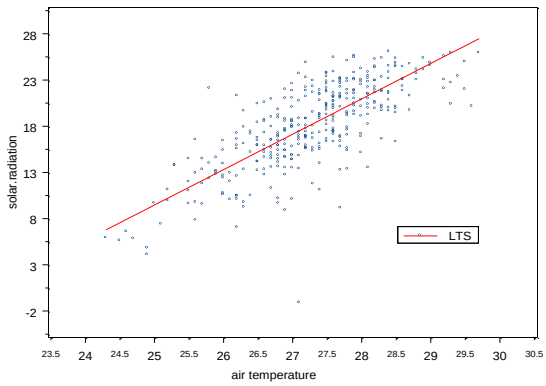

Fig. 7. The scatter plot with the fitted LTS line of air temperature versus solar radiation

The robust multiple R-Squared, $R^2_{LTS} = 0.5993$; means that, about 60 percent of the variability in temperature is accounted for by the straight-line fit to solar radiation. Result shows that, by using the LTS method the coefficient of determination, R^2 value is higher compare to the LS method by 7%.

4) Comparison Results between LS and LTS

The scatter plot with the fitted LS and robust LTS lines of air temperature versus solar radiation is plotted in Fig. 8. From this fig., it was found that the difference fitted between LS and robust LST regression methods are small. LS is compared to highly regard LTS in terms of coefficient of determination, minimum, maximum and total value of estimate to get the best fitted model. One of the effective performance statistics is the coefficient of determination (R^2). The value of R^2_{LS} is 0.5585. Meanwhile, the value R^2_{LTS} is 0.5993. Therefore, the robust LTS regression method is the best fitted model compared to LS method.

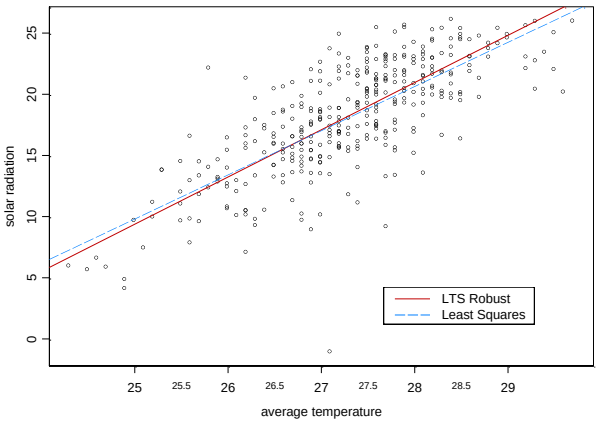

Fig. 8. The scatter plot with the fitted least squares and LTS robust lines of air temperature versus solar radiation

estimation of solar radiation and air temperature compared to the data from weather station *y*. The minimum and maximum data for *y* is -1.11 and 26.06, for LS are 7.267 and 26.761, and the LTS is 6.675 and 27.516, respectively. The min value of solar radiation *y* is -1.11 might be due to rain events or/and higher air mass. The highest air mass somewhat reduces the clear sky data by the absorption along the longer path length. Notice that values such as the mean and median of $\hat{y}_{LS}$ are closed to *y* also similar is the value of $\hat{y}_{LTS}$ as both of the latter. The total value of *y*, $\hat{y}_{LS}$ and $\hat{y}_{LTS}$ are 6592.10, 6592.41 and 6648.64, respectively. Based on analysis it shows that LTS method can also be used to estimate the solar radiation, this method can be used to predict raw data on solar radiation. This means that the estimation by using LTS method is considered as a good method to estimate solar radiation. The reason for this is that the LTS method is not easily influence by the outliers of data set so it can be considered as the best fit for solar radiation model compare to the LS regression method.

TABLE I
COMPARISON OF LS REGRESSION AND LTS ROBUST REGRESSION

Descriptive Measure	Min	Max	Mean	Median	Total
Solar radiation, y	-1.11	26.06	18.06	18.5	6592.1
Estimated Solar radiation using LS	7.267	26.761	18.061	18.458	6592.41
Estimated Solar radiation using LTS	6.675	27.516	18.215	18.639	6648.64

V. CONCLUSION

As a conclusion, LTS regression model is the best fitted model compared to LS regression model mainly because the value of R^2_{LTS} is higher, also, by using the LTS method the coefficient of determination, R^2 value is higher compare to the LS method by 7%. Based on analysis it shows that LTS method can also be used to estimate the solar radiation, as it produces similar data as the weather station's data. The reason for this is that the LTS method is not easily influence by the outliers of data set so it can be considered as the best fit for solar radiation model compare to the LS regression method.

ACKNOWLEDGMENT

The authors wish to thank the Centre Excellence of Renewable Energy (CERE) and School of Electrical System Engineering, University Malaysia Perlis (UniMAP) for the technical and also to the Fundamental Research Grant Scheme 2011 (FRGS) for financial support as well.

REFERENCES

[1] S.Adnan, A.Erol, O.Mehmet, E. Galip Kanit, Solar-energy potential in Turkey, *Journal of Applied Energy*, 80, pp. 367-381, 2005.
[2] D.H.W. Li, T.N.T. Lam, V.W.C. Chu, Relationship between the total solar radiation on tilted surfaces and the sunshine hours in Hong Kong, *Journal of Solar Energy*, 82, pp. 1220–1228, 2008.
[3] T. Markvart, *Solar Electricity* (John Wiley & Sons, LTD., New York, 1994).
[4] S.A. Ahmed, Wind energy as a potential generation source at Ras Benas, Egypt, *Renewable and Sustainable Energy Reviews*, 14, pp. 2167-2173, 2010.
[5] The Albedo Effect, Link http://www.ecocem.environmental & albedo
[6] F. Jiang, Investigation of Solar Energy for Photovoltaic Application in Singapore, *The 8th International Power Engineering Conference, IPEC*, pp. 86 – 89, 2007.
[7] A. Itagaki, H. Okamura, M. Yamada, Preparation of Meteorological Data Set Throughout Japan For Suitable Design of PV Systems, *3rd World Conference on Photovoltaic Energy Conversion*, pp. 2074 – 2077, 2003.
[8] A. Mellit, S.A. Kalogirou, S. Shaari, H. Salhi, A. Hadj Arab, Methodology for predicting sequences of mean monthly clearness index and daily solar radiation data in remote areas: Application for sizing a stand-alone PV system, *Renewable Energy*, 33(2008),p 1570-1590
[9] Y.Shijun, Y.Hongxing, The potential electricity generating capacity of BIPV in Hong Kong, *Photovoltaic Specialists Conference, 1997, Conference Record of the Twenty-Sixth IEEE*, pp. 1345-1348, 1997
[10] I. Daut, M. Irwanto , Y.M. Irwan, N. Gomesh, N.S. Ahmad, Combination of Hargreaves method and linear regression as a new method to estimate solar radiation in Perlis, Northern Malaysia, *Journal of Solar Energy*, 85, pp. 2871–2880, 2011.
[11] S. Ibrahim, I.Daut, Y.M. Irwan, M. Irwanto, N. Gomesh, Z. Farhana, Linear Regression Model in Estimating Solar Radiation in Perlis, *Journal of Energy Procedia* 18 (2012) iii–viii, pp. 1413-1424
[12] Maria Carmela Di Piazza, Antonella Ragusa, Gianpaolo Vitale, Identification of Photovoltaic Array Model Parameters by Robust Linear Regression Methods, *International Conference on Renewable Energies and Power Quality (ICREPQ 09)*, 2009.
[13] Adnan Sozen, Erol Arcaklioglu, Solar potential in Turkey, *Journal of Applied Energy* 80 (2005) 35–45.
[14] M. Yorukoglu, and A.N. Celik, A Critical Review on the Estimation of Daily Global Solar Radiation from Sunshine Duration, *Energy Conversion & Management*, 2006. pp. 2441-2450.
[15] Montgomery, D. C, Peck, E. A, *Introduction to linear regression analysis*, 2nd ed., Wiley. New York, 1992.
[16] Barnett, V. and Lewis, T, *Outliers in Statistical Data* (Wiley and son, New York, 1994).
[17] Belsley, D. A., Kuh, E. and Welsch, R. E. *Regression Diagnostics: Identifying influential data and sources of collinearity* (John Wiley & Sons, New York, 1980).
[18] Rousseeuw. P. J. Least Median of Squares Regression, *Journal of the American Statistical Association*, 79, pp. 388, 1984.