Regression analysis of a photovoltaic (PV) system in FPO

A. El Mouatasim, and Y. Darmane

Citation: AIP Conference Proceedings **2056**, 020008 (2018); doi: 10.1063/1.5084981
View online: https://doi.org/10.1063/1.5084981
View Table of Contents: http://aip.scitation.org/toc/apc/2056/1
Published by the American Institute of Physics

Regression Analysis of a Photovoltaic (PV) System in FPO

A. El Mouatasim[1,a)] and Y. Darmane[2]

[1]*Department of mathematics and computer sciences, Polydisciplinary Faculty of Ouarzazate, University of Ibn Zohr, Morocco .*
[2]*Department of physics and chemist, Polydisciplinary Faculty of Ouarzazate, University of Ibn Zohr, Morocco.*

[a)]Corresponding author: a.elmouatasim@uiz.ac.ma

Abstract. Electrical power output variable from PV systems in outdoor conditions is substantially influenced by climatic input variables such as solar irradiance, wind, dust and temperature. In this work we proposed different types of modeling technique such as simple linear regression, multiple linear regression and non-linear regression for analyzing a PV systems. Then estimates of regression model's parameters should be obtained accordingly, using some optimization methods like gradient method and Matlab function of regression model and machine learning, which relies on minimizing the sum of square of errors.
Regression analysis describes the relationship between a dependent variable and several independent variables. An application on real data set is also provided from PV systems with three technologies polycrystalline, amorphous, and monocrystalline in polydisciplinary faculty of Ourzazate (FPO) - Morocco.
The dependent variable consisted in electrical power, while the independent variables were the following: solar irradiance, ambient temperature and module temperature. A mathematical equation is used to estimate the electrical power.

INTRODUCTION

The major problem related to the photovoltaic PV solar energy and the consequent scientific challenge is that its production greatly depends on the climatic parameters such as solar irradiance, temperature, dust, wind speed and humidity, where the PV plant is installed. Prediction of the PV solar energy production for hours or days ahead can contribute to an efficient and economic use of this resource and can allow to manage the amount of energy obtained by PV plants in order to satisfy the growing demand of it. Furthermore, the increasing development of electrical smart grids has led to the need of knowing in advance the energy production from renewable sources in order to manage the energy flows within the smart grid itself: Forecasting the power output of a PV plant for the next hours or days is necessary for the optimal integration of this production into power systems. There were many scientific investigations in this area carried out in recent years, and different forecasting strategies have been used to achieve the desired goals. In most cases, the studies have focused on the prediction of solar radiation [1], production forecast [2, 3].
Forecasting accurately PV power output is not only based on climatic conditions but large variation of the power output is also observed due to several factors such as solar irradiance, wind, dust, and temperature [4].
Due to the significant effect, numerous studies have been conducted in establishing PV models. Different types of models such as simple linear regression, multiple linear regression [1, 3, 6].
Linear regression analysis can be defined as the process of developing a mathematical model that can be used to predict one variable Y called dependent variable by using another variable or variables X called independent variable. Correlation analysis measures the degree of linear association between two variables.
Error is inherent in data. When data exhibits substantial error rigorous techniques must be used to fit the "best" curve to the data. Otherwise prediction of intermediate values, or the derivatives of values, may yield unsatisfactory results. Visual inspection may be used to fit the "best" line through data points, but this method is very subjective. Some criterion must be devised as a basis for the fit. One criterion would be to derive a curve that minimizes the discrepancy between the data points and the curve. Least-squares regression is one technique for accomplishing this objective. In this work, we have used the data provided from PV systems with three technologies amorphous, monocrystalline and polycrystalline in polydisciplinary faculty of Ourzazate - Morocco. The data includes: solar irradiance, ambient temperature, module temperature, wind and the generated output power. The data is collected every 5 minutes, and is

1st International Congress on Solar Energy Research, Technology and Applications (ICSERTA 2018)
AIP Conf. Proc. 2056, 020008-1–020008-11; https://doi.org/10.1063/1.5084981
Published by AIP Publishing. 978-0-7354-1784-7/$30.00

available from January 1, 2017 up to date Mars 31, 2017.

LINEAR REGRESSION MODELS

Simple linear regression

Regression analysis is a statistical technique for estimating the relationship among variables which have reason and result relation. Main focus of univariate regression is analyses the relationship between a dependent variables $X_1, \ldots, X_n$ and one independent variable Y and formulates the linear relation equation between dependent and independent variable.

The simple linear regression model is the simplest regression model in which we have only one predictor X. This model, which is common in practice, is written as

$$Y_i = b + aX_i + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where

- $Y_i,\ X_i$ are the values of the response and predictor variables in the trial, respectively;
- the unknown parameters: a is called the intercept, and b is the slope of the line;
- ϵ_i are usually assumed to be iid (error) from $N(0, \sigma_\epsilon^2)$ specially for inference purposes (see for instance

Then estimates of simple linear model's parameters should be obtained accordingly, using some method like ℓ_1 regularization [7], the ordinary least squares method, which relies on minimizing the sum of square of errors $\sum \epsilon_i^2$. For the simple linear regression model the ordinary least squares estimations of a and b are

$$\hat{b} = \bar{Y} - \hat{a}\bar{X}$$

and

$$\hat{a} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}.$$

The goodness R^2 of fit is defined as

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}.$$

Matrix form of multiple regression

Regression models with one dependent variable and more than one independent variables are called multiple linear regression.

Multivariate regression analysis model is formulated as in the following:

$$Y = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_n X_p + \epsilon$$

where

- Y is the dependent variable,
- X_i are the independent variables,
- α_i are the parameters,
- ϵ is the error.

The assumptions of multi-linear regression analysis are normal distribution, linearity, freedom extreme values and having no multiple ties between independent variables.

The linear model can be written as

$$Y = X\alpha + \epsilon$$

where

- $Y \in \Re^n$: the vector of observations on the dependent variable,

$$Y = (Y_1, \ldots, Y_n)^T.$$

- $X \in \Re^n \times \Re^{p+1}$: the matrix consisting of a column of ones and p column vectors of the observations on the independent variables,

$$M = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & X_{23} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & \ldots & X_{np} \end{pmatrix}.$$

- $\alpha \in \Re^{p+1}$: the vector of parameters to be estimated,

$$\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)^T.$$

- $\epsilon \in \Re^n$: the vector of random errors,

$$\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T.$$

The vector α is a vector of unknown constants to be estimated from the data by $\hat{\alpha}$.

The normal equations [6] are written as

$$X^T X\hat{\alpha} = X^T Y.$$

If $X^T X$ has an inverse, then the unique solution of normal equations given by

$$\hat{\alpha} = (X^T X)^{-1}(X^T Y).$$

The vector $\hat{Y}$ of estimated means of the dependent variable Y for the values of the independent variables X_1, X_2, ..., X_n in the dataset is computed as

$$\hat{Y} = X\hat{\alpha}.$$

However, to express $\hat{Y}$ as a linear function of Y. Thus,

$$\hat{Y} = [X(X^T X)^{-1} X^T]Y.$$

Correlation

R is known as the correlation coefficient and its value determines the strength and direction of linear association between the two variables under examination. In other words, the value of R will tell us whether there is a relationship between the two variables and how strong that relationship is. If there is a relationship, then the value of R will also indicate whether the value of the dependent variable increases or decreases as the value of the independent variable goes up. The values of the correlation coefficient R near 1 or 1 indicate a strong correlation between the two variables, whereas values of the correlation coefficient near zero indicate no correlation between the two variables.

Dataset of Photovoltaic System in FPO

PV system are installed in FPO as follow:

1. Model 1: connected 8 solar plat of amorphous in series figure 1;
2. Model 2: connected 6 solar plat of monocrystalline in series figure 2;
3. Model 3: connected 6 plat of polycrystalline in series figure 3;
4. model 1, model 2 and model 3 are parallel connected.

The data is collected every day in January 2017 to Mars 2017 in FPO PV system at each 5 minutes for the following measures:

- output power of each models;
- inclined solar irradiance;
- horizontal solar irradiance;
- temperature;
- wind.

FIGURE 1. Amorphous PV module.

FIGURE 2. Monocrystalline PV module.

FIGURE 3. Polycrystalline PV module.

SIMPLE AND MULTIPLE LINEAR REGRESSION OF PV IN FPO

Simple linear regression of the power PV in FPO

We define the following independent variables for PV model in FPO:
X: the output PV power at 12:00-12:05.
Also we defined the following dependent variables:
Y: the output PV power cumulated in day.
We used the following Matlab commend:

```
plotregression(X,Y)
```

The parameters estimated a_1 , a_2 and the coefficient of correlation R is showed in figure 4.
For estimation and test we use the following Matlab commend:

```
lm = fitlm(X,Y)
```

we have the following results:

```
>>
lm =
```

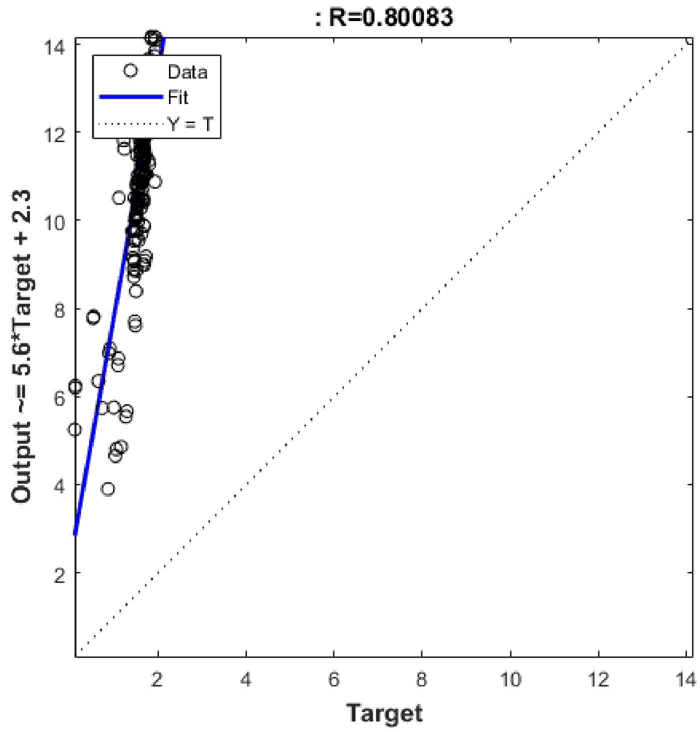

FIGURE 4. SLR fit of PV power in FPO.

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    ______    __________

    (Intercept)     2.2983     0.44014    5.2218    3.9355e-07
    x1              5.5741     0.27368    20.367    1.4422e-53


Number of observations: 234, Error degrees of freedom: 232
Root Mean Squared Error: 1.16
R-squared: 0.641,  Adjusted R-Squared 0.64
F-statistic vs. constant model: 415, p-value = 1.44e-53
```

Multi linear regression of the PV model in FPO

Variable of the PV model

In this section , we use the following data collected every day in January 2017 at 12:00-12:05. We define the following independent variables for PV model in FPO:

- X1: Inclined solar irradiance,

- X2: Horizontal solar irradiance,
- X3: Temperature.

Also we defined the following dependent variables:

- Y1: the cumulated output power of polycrystalline model in day,
- Y2: the cumulated output power of amorphous model in day,
- Y3: the cumulated output power of monocrystalline model in day.

Analysis of PV system in FPO

In this section, we studies the power total of PV system in FPO for three month in 2017.
Then $Y = Y_1 + Y_2 + Y_3$. Create linear regression model by fitlm Matlab function.
We examine the quality of the fitted model, consult an ANOVA table.

	SumSq	DF	MeanSq	F	pValue
	------	--	------	-------	---------
x1	76.423	1	76.423	8.6776	0.0043046
x2	64.696	1	64.696	7.346	0.0083491
x3	1.5334	1	1.5334	0.17412	0.67769

Depending on our goals, consider removing X_3 from the model.

```
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate        SE          tStat       pValue
                   --------     ---------     ------     ----------

    (Intercept)      8.9778        1.8705     4.7998     7.9106e-06
    x1             0.011706     0.0039097     2.9941      0.0037271
    x2              0.01696     0.0044131     3.8431     0.00025248


Number of observations: 78, Error degrees of freedom: 75
Root Mean Squared Error: 2.95
R-squared: 0.71,  Adjusted R-Squared 0.702
F-statistic vs. constant model: 91.7, p-value = 7.06e-21
```

Diagnostic Plots

We use diagnostic plots for identify outliers, and see other problems in our model or fit. We show in figure 5, there are a few points with high leverage and one point with large Cooks distance. Identify it and remove it from the model. You can use the Data Cursor to click the outlier and identify it, or identify it programmatically:

```
[~,larg] = max(lm.Diagnostics.CooksDistance);
lm2 = fitlm(X,Y,'Exclude',larg);
```

Residuals Model Quality for Training Data

There are several residual plots that can used for discover errors, outliers, or correlations in the model or data. The simplest residual plots are the default histogram plot, which shows the range of the residuals and their frequencies, and the probability plot, which shows how the distribution of the residuals compares to a normal distribution with matched variance.

Examine the residuals:

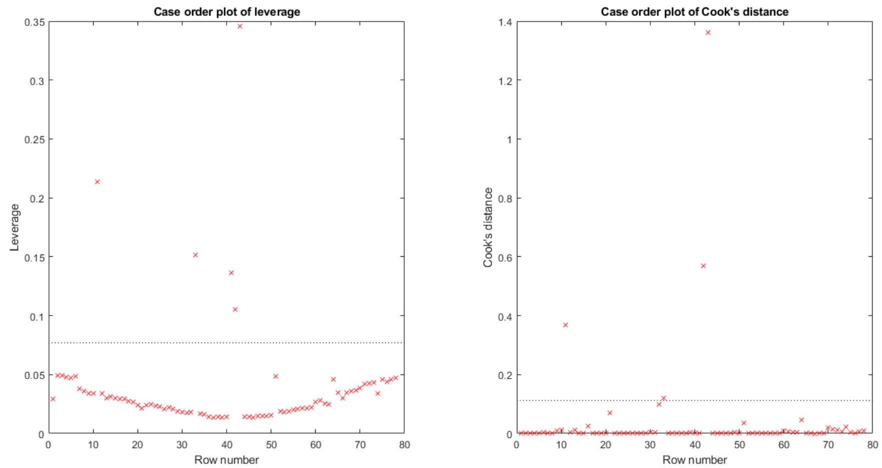

FIGURE 5. Diagnostic of PV module.

```
plotResiduals(lm)
```

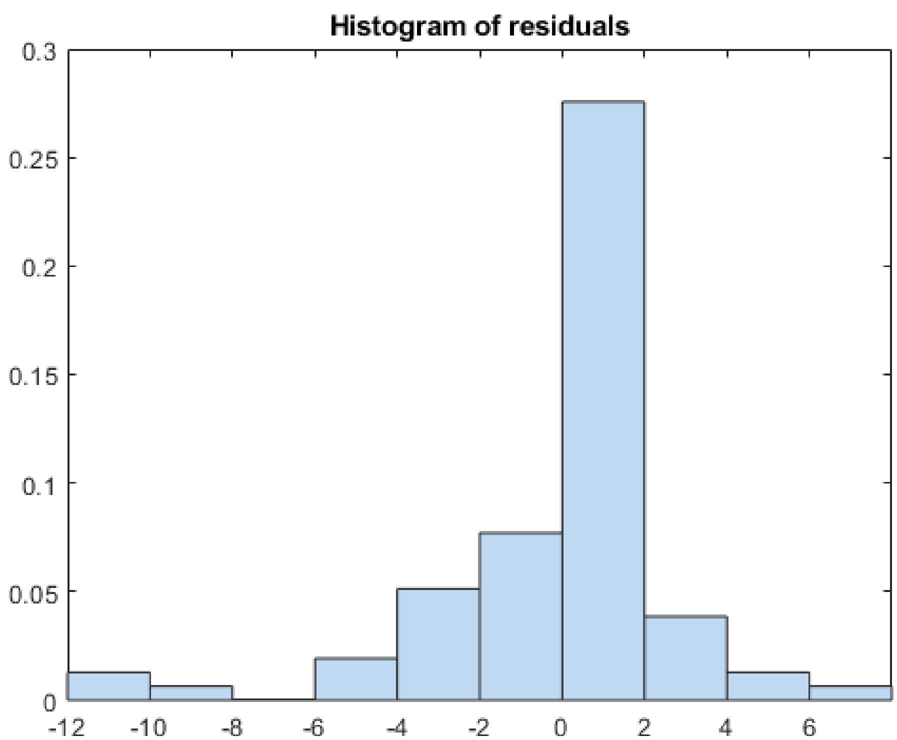

FIGURE 6. Histogram before.

The observations less than -8 are potential outliers.

```
plotResiduals(lm,'probability')
```

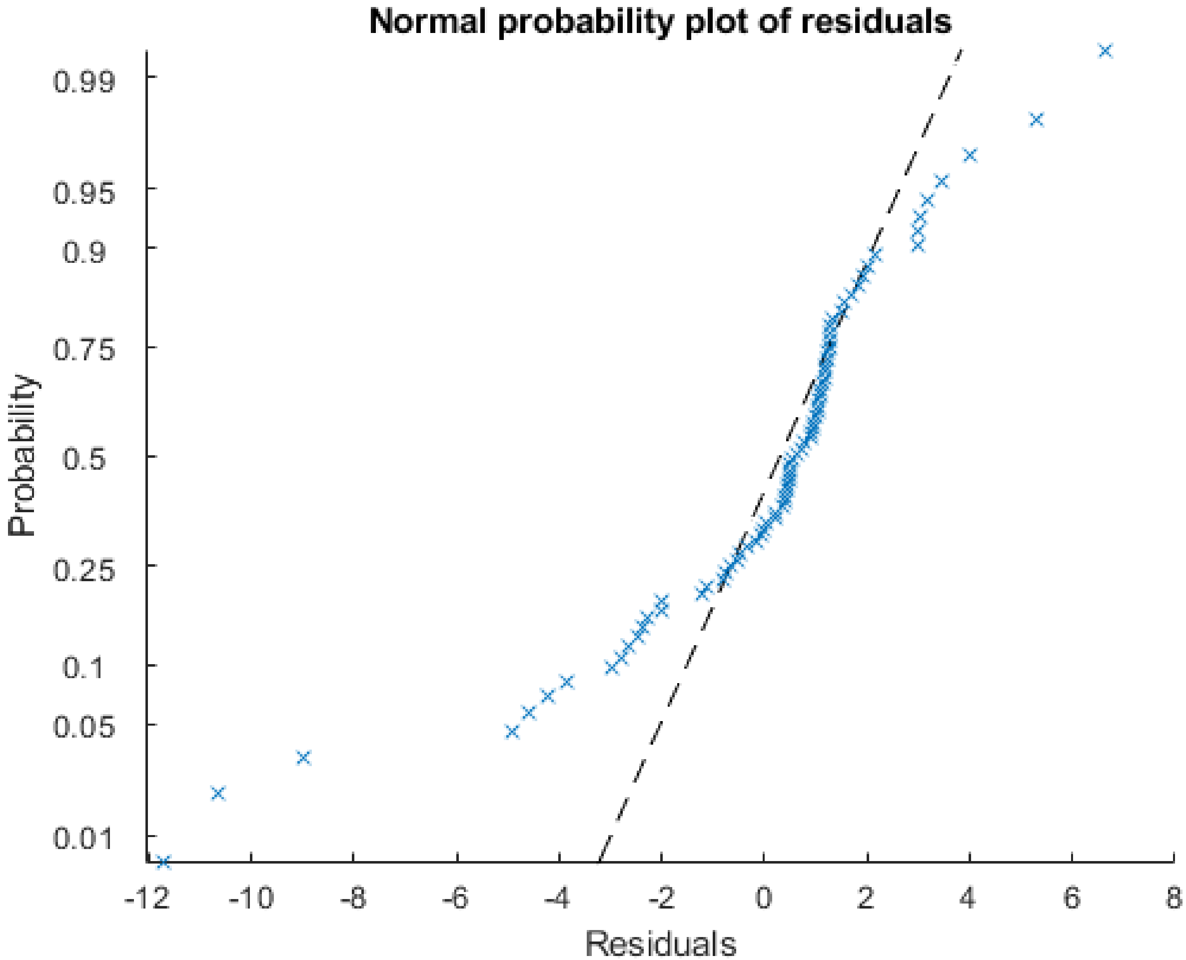

FIGURE 7. Probability plot.

The three potential outliers appear on this plot as well. Otherwise, the probability plot seems reasonably straight, meaning a reasonable fit to normally distributed residuals.

You can identify the three outliers and remove them from the data:

```
outl = find(lm.Residuals.Raw<−8 );
```

To remove the outliers, use the exclude name-value pair:

```
lm2 = fitlm(X,Y,'Exclude',outl);
```

Examine a residuals plot of lm2:

```
plotResiduals(lm2)
```

The new residuals plot looks fairly symmetric, without obvious problems. However, there might be some serial correlation among the residuals. Create a new plot to see if such an effect exists.

Plots to Understand Predictor Effects

Examine a slice plot of the responses. This displays the effect of each predictor separately.

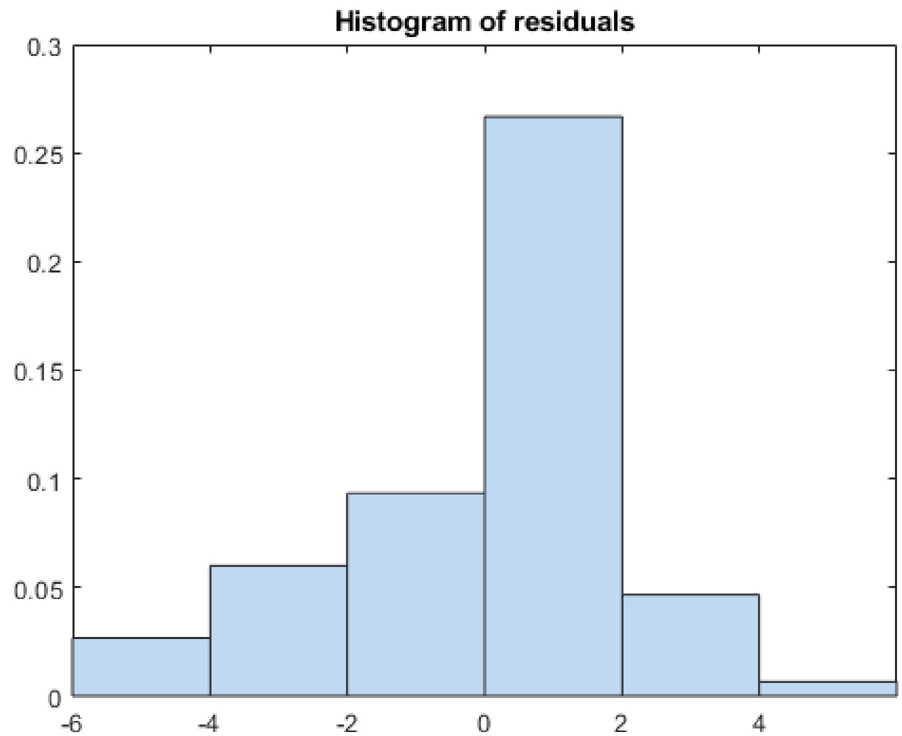

FIGURE 8. Histogram after.

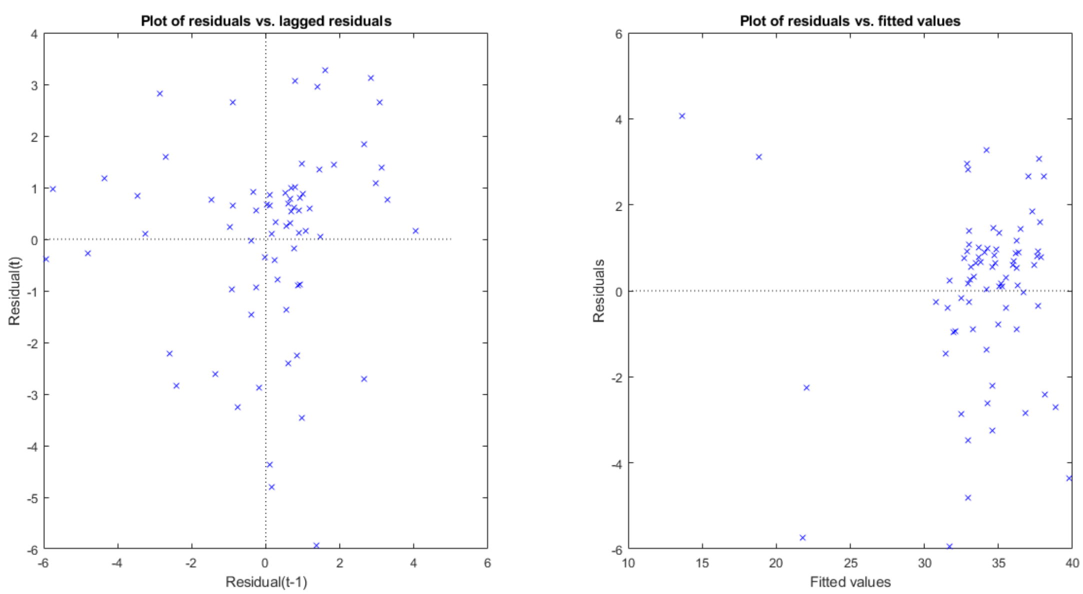

FIGURE 9. Residuals plot

```
tbl = table(x1,x2,Y,'VariableNames',{'InclinedIrrad','HorizontalIrrad','Power'});
lm = fitlm(tbl,'Power~InclinedIrrad+HorizontalIrrad');
 plotSlice(lm)
```

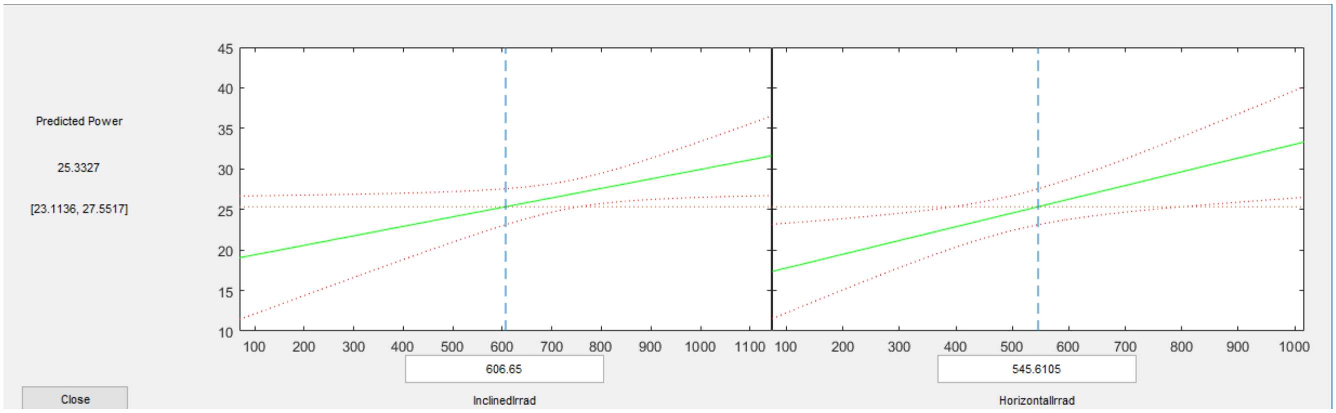

FIGURE 10. Prediction Slice plots

Use an effects plot to show another view of the effect of predictors on the response.

```
plotEffects(lm)
```

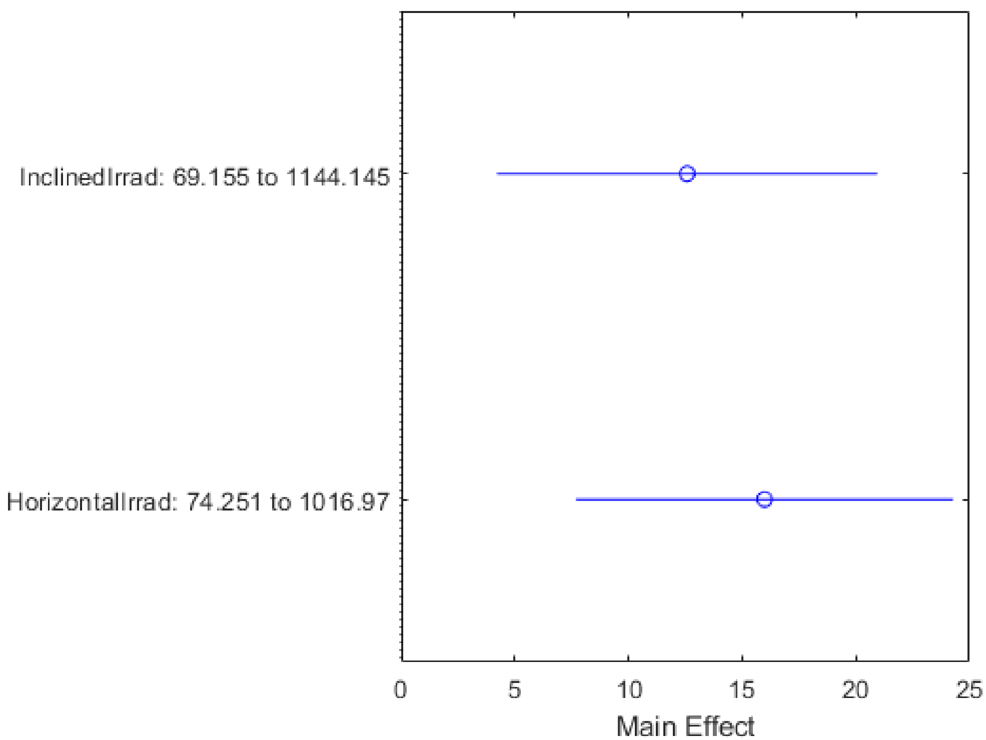

FIGURE 11. Effects plot

This plot shows that changing Inclined irradiation from about 70 to 1144 lowers Power by about 12 (the location of the upper blue circle). It also shows that changing the horizontal irradiation from 74 to 1017 raises Power by about 16 (the lower blue circle). The horizontal blue lines represent confidence intervals for these predictions. The predictions come from averaging over one predictor as the other is changed. In cases such as this, where the two predictors are correlated, be careful when interpreting the results.

CONCLUSION

According to results above, the simple and multiple line-ar regression models can be used to estimate the power of polycrystalline PV module, amorphous PV module and monocrystalline PV module in FPO, Morocco. The relationship between solar irradiance, temperature and output power is linear. The linear correlation coefficient value and the value of R^2 in SLR model and MLR model are higher.
Future work is statistical and neural network comparison of out power PV of polycrystalline, amorphous and monocrystalline in FPO.

REFERENCES

[1] S. Ibrahim, I. Daut, Y.M. Irwan, M. Irwanto, N. Gomesh and Z. Farhana, “Linear regression model in estimating solar radiation in Perlis,” Energy Procedia, vol. 18, pp.1402-1412,(2012).

[2] R. De Leone, M. Pietrini and A. Giovannelli, “Photovoltaic energy production forecast using support vector regression,” Neural Computing & Application, to appear, Volume 26, Issue 8, pp. 1955-1962, (November 2015).

[3] G. Wang, Y. Su and L. Shu, “One-day-ahead daily power forecasting of photovoltaic systems based on partial function linear regression models,” Renewable Energy, vol. 96, pp. 469-478, (2016).

[4] H. Ramenah, P. Casin, M. Ba, M. Benne and C. Tanougast, “Accurate determination of parameters relationship for photovoltaic power output by augmented dickey fuller test and engle granger method,” AIMS Energy, vol. 6, no. 1, pp: 19-48, (2018).

[5] V. E. Sunday, O. Simeon and U.M. Anthony, “Multiple linear regression photovoltaic cell temperature model for PVSyst simulation software,” International Journal of Theoretical and Applied Mathematics, vol. 2, no. 2, pp. 140-143, (2016).

[6] A. El Mouatasim, “Simple and multiple linear regression of verbs in Quran,” American Journal of Computation Mathematics, vol. 8, pp. 68-77, (2018).

[7] A. El Mouatasim and M. Wakrim, “Control subgradient algorithm for image ℓ_1 regularization,” Journal of Signal and Video Processing (SViP), vol. 9, pp. 275-283, (2015).