Robust Modeling for Factors Affecting in Relative Humidity in Basra Governorate

<table>
<thead>
<tr>
<th>Nazik J. Sadik</th>
<th>Nabaa N. Mahdi</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:dr.nazik@coadec.uobaghdad.edu.iq">dr.nazik@coadec.uobaghdad.edu.iq</a></td>
<td><a href="mailto:nabaanaemmahdi@uomustansiriya.edu.iq">nabaanaemmahdi@uomustansiriya.edu.iq</a></td>
</tr>
<tr>
<td>Department of Statistics - College of Administration and Economic - Baghdad University, Baghdad, Iraq</td>
<td>Department of Statistics - College of Administration and Economic - Mustansiriya University, Baghdad, Iraq</td>
</tr>
</tbody>
</table>

Husam A. Rasheed

husamstat@uomustansiriya.edu.iq

Department of Statistics - College of Administration and Economic - Mustansiriya University, Baghdad, Iraq

**Article Information**

**Abstract**

Basra Governorate is considered among the Iraqi governorates, with high levels of relative humidity sometimes, which makes the weather uncomfortable during the year. It was necessary to building in this research a robust regression model to study the effect of some factors such as the maximum and minimum temperatures, the amount of rainfall, atmospheric pressure, and the average amount of dust falling. When building linear regression models, many problems occur in which failure some of the hypotheses model. In this research, least squares estimates were compared with robust and found the preference for the method of S.

**1. Introduction**

One of the most important statistical methods is modeling different phenomena with models that are closer to reality, and these models must have efficient capabilities that correctly express the studied phenomenon so that we can predict and develop future plans and build strategies. Either they are probabilistic, such as time series models, Markov chains, and reliability models, or they are causal, meaning that the construction of these models is a cause and effect of the cause, and this is what is called regression models that are built on a relationship between the explanatory variables or what is called explanatory and the dependent variable or the response variable.

The German mathematician (Carl Friedrich Gauss) introduced the Ordinary Least Squares method which is the best unbiased linear estimation (BLUE) under the basic assumptions, Where the random errors are independent and normally distributed with the least variance among all other linear estimators, and the presence of any problem such as the presence of outliers or extreme values would make them less accurate for the deviation of the distribution and make it twisted, and
this makes these estimators inefficient, so we resort to other estimation methods that are more flexible, which are the robust estimators. Due to the existence of such variables in our practical life, this problem has received great attention, and the reason is the realization that the use of traditional methods in estimating parameters leads to the emergence of the problem and the data is not fully subject to the assumptions made. It leads to incorrect and unsatisfactory results in estimation and prediction.

2. Problem of Search
Climate change in the world in general and in Iraq in particular is caused by the presence of observations or data that contain abnormal values, or that the distribution of errors does not follow the normal distribution, and thus there will be a breach in the assumptions of the model that was built for the purpose of predictive estimation, and therefore correct data must be dealt with and analyzed in order to obtain forecasts or estimates.

3. Research objective
The research aims to find the best robust estimator for data that penetrates the method of least squares in finding the estimators of the model under study. These methods are the method of maximum likelihood or what is called the (M) method, the (MM) method, and the (S) method. A comparison was made between the three methods using the mean error standard (MSE).

4. Robust Regression [1,2,3]
The robust regression is used to estimate the parameters of the regression model when the data suffers from the problem of the presence of outliers or extreme values, so the traditional estimation methods become useless, and we resort to estimating the model parameters by alternative methods that are resistant to addressing these problems. The efficiency of the traditional methods is equivalent to the usual least squares method, and it is more efficient than it in the event of a problem.

Robust Method
There are many robust methods that have been suggested by researchers, and among those methods we will present the following:

M Method
This method is considered one of the most widely used methods of robust regression. It was proposed by the scientist Hooper in 1973. It is not affected by outliers, in addition to its ease of theoretical use. It is built on the basis of:

1- Minimize the objective function instead of decreasing the sum of squares of the error of the objective function [1, 4], as in the following formula:

\[
\min \sum_{i=1}^{n} \eta \left( \frac{e_i}{s} \right)
\]

2- The error can be found through the following formula:

\[
e_i = y_i - x_i \beta
\]

3- Find the standard deviation according to the following formula:

\[s = 1.483MAD\]

Where, the median of the absolute deviation is calculated according to the formula:

\[MAD = \text{median} |e_i - \text{median}(e_i)|\]

4- Obtain the parameters of the model, it is by taking the partial derivatives with respect to the parameters and equating them to zero, and we get this equation:

\[\sum_{i=1}^{n} \psi \left( \frac{y_i - x_i \beta}{s} \right) x_i = 0\]
The selection of the function $\psi$ depends on the preference of the amount of weight to determine the outliers, and this function is characterized by the following characteristics:

\[ p(e) \geq 0, \quad p(0) = 0, \quad p(e) = p(-e), \quad p(e_i) \geq p(e_i') \text{ for } |e_i| \geq |e_i'| \]

This means that the probability of the zero value of the error is equal to zero, and that the positive and negative values have the same probability. The probability of the original function of error is greater than the probability of the derivative of the error, and by defining the weight function, the equation can be solved: [5, 7]

\[
w(e_i) = \frac{\psi\left(\frac{y_i - x'_i\hat{\beta}}{s}\right)}{\psi\left(\frac{y_i - x'_i\hat{\beta}}{s}\right)}
\]

Equation (5) can be written in the following form:

\[
\sum_{i=1}^{n} w\left(\frac{y_i - x'_i\hat{\beta}}{s}\right)x_i = 0
\]

We can write the above formula in matrix form:

\[
X'WX\hat{\beta}_M = X'WY
\]

The estimator for M can be obtained according to the following formula:

\[
\hat{\beta}_M = (X'WX)^{-1}X'WY
\]

Where $W$ is the matrix of weights

**MM Method [4]**

This method depends on the use of the robust (S) estimation by minimizing the residual measure of the (M) method, so it is considered a follower of the (M) estimation method. It has high efficiency in the case of normal distribution of errors with a high breaking point. This estimator can be obtained by following the following steps:[9,8]

1. Determine an initial estimator with a high breakdown point, and it is not necessarily efficient, and through it we calculate the initial residuals according to the following formula:

\[
r_i(\hat{\beta}_S) = y_i - x'_i\hat{\beta}_S, \quad 1 < i < n
\]

2. Calculate the estimator (M) of the (measurement (Sn)) for the residuals depending on the estimated (M) equation for the measurement parameter and according to the following formula:

\[
\hat{\beta}_M = \left(\sum_{i=1}^{n} P\left(r_i(\hat{\beta}_S)/s\right)\right)/n = 1/2
\]

Whereas:

\[
S_n = S\left(r_1(\hat{\beta}_S), \ldots, r_n(\hat{\beta}_S)\right)
\]

3. The estimator MM can be defined as the estimator M of $\beta$ using a function (re-descending)

\[
\psi(u) = \frac{\partial p_1(u)}{\partial u}
\]

4. By solving the following equation, we will get the MM estimator.

\[
\sum_{i=1}^{n} X_{ij} \psi_1\left(\frac{y_i - x'_i\beta}{s_n}\right) = 0
\]

And the measurement estimate can be obtained from the second step, as:

\[
r_i(\hat{\beta}) = y_i - x'_i\hat{\beta}
\]
One of the disadvantages of the M method is the loss of the data distribution as well as the probability function of the data due to the use of the median only as a weight value. The S method is based on the estimation of the residuals of the M method, and the standard deviation of the residuals is used in order to get rid of the weaknesses. The S estimator can be defined according to the following formula \[^{[8][10]}\):

$$\hat{\beta}_S = \min_\beta \sum_{i=1}^{n} P \left( \frac{y_i - \sum_{j=0}^{k} x_{ij} \beta_j}{\hat{\sigma}_S} \right)$$

and is determined \(\hat{\sigma}_S\) minimum robust scale estimator according to the following equation:

$$\min_\beta \sum_{i=1}^{n} P \left( \frac{y_i - \sum_{j=0}^{k} x_{ij} \beta_j}{\hat{\sigma}_S} \right)$$

\(\hat{\sigma}_S\) can be calculated according to the following formula:

$$\hat{\sigma}_S = \left( \frac{1}{0.199} \sum_{i=1}^{n} w_i e_i^2 \right)$$

As for how to obtain the initial estimate, we depend on equation (4) and divide by (0.6745). Thus, the solution can be obtained by substituting the parameters according to the following:

$$\sum_{i=1}^{n} x_i \psi \left( \frac{y_i - \sum_{j=0}^{k} x_{ij} \beta_j}{\hat{\sigma}_S} \right) = 0 \quad j = 0, 1, \ldots, k$$

Since \(\psi\) is a function derived from \(P\):

$$\psi(u_i) = \begin{cases} u_i(1 - (u_i/c)^2) & |u_i| \leq c \\ 0 & |u_i| > c \end{cases}$$

5. The Practical side

Data on the weather conditions of Basra Governorate for the period from (1-2015 to 12-2020) were collected from the Central Statistical Organization website, where the independent variables represent, Monthly Average Precipitation (MAP), General Average Atmospheric Pressure (GAP), and General Average Normal Temperature (GANOT), while the dependent variable represents the General Average Relative Humidity (GARH).

![Figure 1: General Average relative humidity (2015-2020)](image-url)
Fig. (2): Monthly Average precipitation (millimeters) (2015-2020)

Fig. (3): General Average atmospheric pressure (millibars) (2015-2020)

Fig. (4): General Average normal temperatures (2015-2020)

- Testing outliers in the data
  The data was tested using the box-plot method, and it was proved that there are abnormal values in the model, as shown in the following figures:
Robust Modeling for Factors Affecting in Relative Humidity in Basra Governorate

Nazik J. Sadik; Nabaa N. Mahdi and Husam A. Rasheed

Fig. (5): box-plot General Average normal temperatures

Fig. (6): box-plot of Monthly Average precipitation (millimeters)

Fig. (7): box-plot General Average atmospheric pressure (millibars)

Fig. (8): box-plot General Average relative humidity

• Results of using the robust M method the weight function (Bisquare) and the function (Huber) were used, and the results were as shown in Table (1).

Table (1): estimations of the robust M method

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.314595</td>
<td>0.114946</td>
<td>2.736894</td>
<td>0.0062</td>
</tr>
<tr>
<td>GAP</td>
<td>0.076080</td>
<td>0.004518</td>
<td>16.83982</td>
<td>0.0000</td>
</tr>
<tr>
<td>GANOT</td>
<td>-1.153488</td>
<td>0.133406</td>
<td>-8.646418</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

We note from the above table that there is a direct effect of each of the precipitation (MAP) and atmospheric pressure (GAP) on the general average relative humidity (GARH), while there was an inverse effect of temperature (GANOT) on the relative humidity (GARH), i.e. the greater (GANOT) increased, so (GARH) decreased, for Basra Governorate, where we note from the value of R2, that the explanatory variables explain 65.9% of the dependent variable (GARH).

• Results of using the robust S method

Through the application of the robust S method and as shown in Table (2), the results showed that there is a significant effect of the studied variables on the variable (GARH) of Basra Governorate, where we notice that the higher the MAP, GAP and the lower the GANOT value, the higher the GARH value of Basra Governorate. We notice from the value of R2 that the independent variables explain 91.6% of the effect of the dependent variable, GARH.

Table (2): Parameter of the robust S method

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.609024</td>
<td>0.046005</td>
<td>13.23821</td>
<td>0.0000</td>
</tr>
<tr>
<td>GAP</td>
<td>0.065559</td>
<td>0.001808</td>
<td>36.25653</td>
<td>0.0000</td>
</tr>
<tr>
<td>GANOT</td>
<td>-0.827979</td>
<td>0.053393</td>
<td>-15.50713</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

• Results of the robust MM method The method proved a significant effect of the study variables on the relative humidity (GARH) in Basra Governorate, and the results were similar to the two robust
methods (M, S), that is, there is a direct effect of the two variables (MAP, GAP) and an inverse effect of the variable (GANOT), and the value of $R^2$ shows, That all the independent variables explain 66.4% of the dependent variable, as shown in Table (3).

**Table (3): Estimations of the robust (MM) method**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.389142</td>
<td>0.102638</td>
<td>3.791419</td>
<td>0.0001</td>
</tr>
<tr>
<td>GAP</td>
<td>0.073504</td>
<td>0.004034</td>
<td>18.22075</td>
<td>0.0000</td>
</tr>
<tr>
<td>GANOT</td>
<td>-1.073507</td>
<td>0.119121</td>
<td>-9.011886</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Accordingly, we can choose the best way to represent the impeccable linear regression model by choosing the largest value of the coefficient of determination $R^2$, where the S method was the best way to represent the model, because it has the largest coefficient of determination $R^2$, as shown in Table (4).

**Table (4): Comparison criteria**

<table>
<thead>
<tr>
<th>Methods</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>65.9%</td>
</tr>
<tr>
<td>S</td>
<td>91.6%</td>
</tr>
<tr>
<td>MM</td>
<td>66.4%</td>
</tr>
</tbody>
</table>

6. Conclusions and recommendations

1. There is a strong linear regression relationship between MAP, GAP and the dependent variable GARH.
2. There is a strong and inverse linear regression relationship between the variable GANOT and the dependent variable GARH in Basra Governorate.
3. The best way to estimate the model in the presence of outliers is the robust S method.
4. We recommend studying other variables that affect the relative humidity (GARH) of Basra Governorate.
5. We recommend resorting to the use of renewable energy (solar energy) in the generation of electrical energy to address the increase in humidity in Basra Governorate.

References


[3] Hassan, Tara Ahmed and Reda, Mahdi Saber, "Using robust regression to find the most appropriate model to represent meteorological data in the city of Erbil during the period (1998-2010)", Journal of the College of Administration and Economics, 34th year, Issue (89), 2011.


المستخلص

 تعتبر محافظة البصرة من المحافظات العراقية ذات معدلات الرطوبة النسبية العالية، مما يجعل الطقس في بعض الأحيان غير مريح خلال العام. في هذا البحث بناءً على نموذج احترافي حصصيين لدراسة تأثير بعض العوامل مثل درجات الحرارة العظمى والصغرى، وكمية الأمطار، والضغط الجوي، وتمتستخدم نسبة الغبار المتساقط عند بناء نماذج الإحصاء الخطي. تحدث العديد من المشاكل التي تؤدي إلى فشل بعض نماذج الفرضيات. في هذا البحث تم مقارنة تدريبات المربعات الصغرى مع الحسبنة ووجدت الأفضلية لطريقة.S

الكلمات المفتاحية:
طققية، المربعات الصغرى، الرطوبة النسبية

المراسلة:
نازك جعفر صادق
dr.nazik@coadec.uobaghdad.edu.iq
https://doi.org/10.55562/jrucs.v54i1.618