Panel Data Analysis and Modeling for Wheat Production in Kurdistan Region -Iraq

Assist.Prof. Dr.Nawzad Mohammad Ahmed

Admin.& Eco, Sulaimani University

Heshu Othman Mahmud

Admin.& Eco, Sulaimani University

تحليل ونمذجة البيانات الطولية لانتاج القمح في أقليم كردستان – العراق

أ.م.د. نوزاد محمد احمد هيشو عثمان فقى محمود

كلية الادارة والاقتصاد/ جامعة السليمانية

تاريخ استلام البحث 2019/3/25 تاريخ قبول النشر 2019/5/8

Abstract:

In this study it is required to estimate the economic growth function for wheat production, balanced panel data models was used to reach the main goal to predict wheat production depending on the area in acres and the precipitation in milliliter as explanatory variables the sample of the study contains (60) observations for the three variables explained later in Sulaimani, Hawler and Duhok cities for period(1992-2011).The pooled regression, fixed effect and random effect models have been run for the data under study and the postulated models compared depending on coefficient determination (R^2) and Akaiki information criterion (AIC)and Hausman test to select the more adequate model that has a best performance that represents the phenomena of wheat production. According to the measures of ($R^2$and AIC) the best model have been detected which is fixed effect model with (R^2=%82, AIC=0.63).

Key words: Balanced Panel data, pooled regression model, fixed effect model, Random effect model, Hausman test.

المستخلص:

نهدف في هذه الدراسة الى تقدير دالة النمو الاقتصادي لأنتاج الحنطى بأستخدام بيانات طولية المتوازنة للوصول الى القيم التقديرية لأنتاج الحنطى بالاعتماد على المساحة المزروعة بالدونم مع هطول الامطار بالميليلتر كمتغيرين مستقلين. الدراسة تشمل عينة بحجم (60) مشاهدة للمتغيرات الثلاثة المدروسة في المدن (السليمانية، هولير، دهوك) للفترة الزمنية (1992−2011) ولذلك تقديم وتوضيح كيفية الاختبار والتوافق بين النماذج الثلاثة نموذج الانحدار التجميعي ونموذج التأثيرات الثابتة مع النموذج التأثيرات العشوائية وتم في ذلك الاعتماد على معيار معامل التحديد ومعيار اكايكي للمعلومات واختبار هاوسمان لاختبار واختيار أفضل النموذج. بالاعتماد على قيم (R^2 و AIC) تم تحديد افضل نموذج حيث كانت نموذج التأثير الثابت مع (R^2=%82, AIC=0.63).

الكلمات الدالة: بيانات بانل المتوازنة، نموذج الانحدار التجميعي، نموذج التأثيرات الثابتة، نموذج التأثيرات العشوائية، أختبار هاوسمان.

Chapter One

1-1 Introduction:

There is no doubt that agriculture is the most important sector from the economic sector and it has an effect for the other sectors the development of the agriculture sector doesn't have an effect for the insurance of local foods itself, but it is very significant for reducing jobless and reanimate many other sectors in the country. In the old centuries Kurdistan was a place for agriculture because of having log cabins for producing agriculture such as, land, glass houses, rain, water, good weather if they are used properly and significantly, we can deliver Kurdistan region to the extent of self-sufficiency and access to (income) through the exploitation of agriculture properly and the use of modern and modern methods and politics, because agriculture is the basics of all civilizations. Wheat is one the most important of the product of agriculture that has a big effect on economic that is why all the developed countries trying to find a bazaars for their wheat product and find the alternative for that product. Kurdistan in the ancient countries depends on wheat and it was a food for that places that they product it. Charmu zone is one of the most ancient places that they found wheat in it. Now generally agriculture is in ready for the economic competition in Kurdistan region it will form from (%10-%15) it will affect (%10-%25) of economic of Kurdistan of the indemnification of food such as (cereal, vegetables, fruit, potatoes, chicken, meat with the rotation of the capilcalized of local and appirence of the jobs in the local of the country with the less of governments support for that subject. The world wants wheat a lot USA and Russia they are computing for finding bazaars for their wheat in the last 15 years ago wheat had medium effect on the economics of Kurdistan. In addition, the panel models have recently gained considerable interest in economic studies, especially since they take into account the impact of time change and the effect of the difference between the units in the sample data of the study. Panel data models or longitudinal data or combined time-series/cross-section data are terms used in econometrics and statistics to denote data sets which contain repeated observations on a selection of variables from as set of observation units. Panel data usually gives the researcher a large number of data points, increasing the degrees of freedom and reducing the collinear among explanatory variables–hence improving the efficiency of econometric estimations. More precise, panel data allow a researcher to analyze a number of important economic questions that cannot be addressed using cross-sectional or time-series data sets (cheng Hsiao-2003). Panel data models come in three main forms: First, pooled regression model (PRM) Which is one of the simplest longitudinal data models where assume that the coefficient remain constant across time and cross – section, and second fixed effect model which one way to take in to the heterogeneity among individuals is to allow each individuals to have his own intercept that is mean each individuals intercept is differ from the intercepts of other individuals and the intercepts doesn't vary over time which means time invariant the fixed model also called least square dummy variable, third we have random effect model also called error components model (ECM) is an appropriate specification if we are drawing N individuals randomly from a large population where (ϵ_i) is a random error term distributed normal with mean zero and variance ($\sigma^2{}_{\epsilon}$) .

1-2 Aim of the study:

Estimating and selecting the more adequate model that has a best performance to represents the behaviors of the data under consideration to predict wheat production and through the postulated model, and to compare the difference of individuals in wheat production to find out that wheat production is affected by location in this study.

1-3 Hypothesis of the study:

1- The appropriate and best model is fixed effect model.
2- Rainfall and agriculture area are most effectible factors on the wheat production.

Chapter Two

Methodology

2-1 Panel data models:

Panel data models are study the same group of entities (individuals, firms, states, countries, etc) over time. There are several advantages of using longitudinal data compared with either purely cross-sectional or purely time-series data (Badi baltagi-2005) as it is shown below in seven points:

-Controlling for individual heterogeneity: Panel data suggests that individuals, firms, states or countries are heterogeneous. Which time-series and cross-section studies are not controlling this heterogeneity that run the risk of obtaining biased results.

-Panel data give more informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency. Time-series studies are plagued with multicollinearity. With additional, more informative data one can produce more reliable parameter estimates.

-Panel data is adequate and able to study the dynamics of adjustment. Cross-sectional distributions that look relatively stable hide a multitude of changes.

-Panel data are also well suited to study the duration of economic states like unemployment and poverty, and if these panels are long enough, they can shed light on the speed of adjustments to economic policy changes.

-Panel data is better to identify and measuring the effects that are simply not detectable in pure cross-section or pure time-series data.

-Panel data models allow us to construct and test more complicated behavioral models than purely cross-section or time-series data.

-Micro panel data gathered on individuals, firms and households may be more accurately measured than similar variables measured at the macro level Biases resulting from aggregation over firms or individuals may be reduced or eliminated.

Suppose we have (N) cross sectional observations measured in (T) period of time, then the model for such data should be shown as follow:

$$Y_{it} = \beta_{0(i)} + \sum_{j=1}^{k} \beta_j X_{j(it)} + \epsilon_{it} \qquad (i = 1,2,\dots,N \;\; t = 1,2,\dots,t) \qquad (1)$$

Where Y_{it} represents the (i^{th}) observed value of response variable at (t) period of time.

$\beta_{0(i)}$ Represent the cross value in observation i.

β_j Represents the slope of the regression line.

$X_{j(it)}$ Represent the (j^{th}) explanatory value in (i^{th}) observation at (t^{th}) period of time.

ϵ_{it} Represent the error term in (i^{th}) observation at (t^{th}) period of time.

The panel data models come in three main forms: pooled regression model fixed effect model and random effect model.

2-2-1 pooled regression model:

In this type of model the parameters are estimated under hypothesis of time has no effect then we get this below form equation (1):

$$Y_{it} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{j(it)} + \epsilon_{it} \qquad (i = 1,2, \dots, N \;\; t = 1,2, \dots, t) \qquad (2)$$

That tells OLS can be used to estimate the parameters of the above model when:

$$E(\epsilon_{it}) = 0 \, , v(\epsilon_{it}) = {\sigma_\epsilon}^2$$

After organizing the values of response and explanatory's variable the sample will be of size (N*T).

2-2-2 Fixed effect model:

The aim of using this type of model is to detect that the behaviour of every cross sectional data set alone stays the same or not, that can be done depending on intercepts of each cross sectional data set, in another word it is like to have a heterogeneity in variance among the cross sectional data sets under study, then the model is written as follow:

$$Y_{it} = \beta_{0(i)} + \sum_{j=1}^{k} \beta_j X_{j(it)} + \epsilon_{it} \qquad (i = 1,2, \dots, N \;\; t = 1,2, \dots, t) \qquad (3)$$

Where $(\epsilon_{it}) = 0 \, , v(\epsilon_{it}) = {\sigma_\epsilon}^2$

The meaning behind the fixed effect term is that (β_0) for all cross sectional data sets is not change through time, but the change would be in totals of the cross sectional data sets .

To estimate the parameters of equation (3) usually dummy variable least square method is used to avoid perfect multicolinearity, after adding dummy variables the model is become as below form:

$$Y_{it} = \alpha_1 + \sum_{d=2}^{N} \alpha_d \, D_d + \sum_{j=1}^{k} \beta_j \, X_{j(it)} + \epsilon_{it} \quad (i = 1,2, \dots, N \;\; t = 1,2, \dots, t) \quad (4)$$

The model after removing (α_1) is become on the below form:

$$Y_{it} = \sum_{d=2}^{N} \alpha_d \, D_d + \sum_{j=1}^{k} \beta_j \, X_{j(it)} + \epsilon_{it} \qquad (i = 1,2, \dots, N \;\; t = 1,2, \dots, t) \quad (5)$$

We can provide a test to find out if the fixed effect model is better than the pooled model .Since the pooled model neglects the heterogeneity effects that are explicitly taken to in account in the fixed effect model the pooled model is restricted version of the fixed effects model therefore we can use the restricted F test.

$$F = \frac{({R_{unr}}^2 - {R_r}^2)/m}{(1-{R_{unr}}^2)/(n-k)} \qquad (6)$$

Where (R^2_{unr}) and (R^2_r) are unrestricted and restricted coefficients of determination, **(m)** is the number of parameters omitted from the restricted model, **(n)** is the number of observations in the sample and **(k)** is the number of parameters estimated in the unrestricted regression.

2-2-3 Random effect model:

In fixed effect model $\epsilon_{it} \sim N(0, {\sigma_\epsilon}^2)$, the estimators of parameters should be unbiased also the data set should be homogenous and the postulated model is not suffer from autocorrelation problem. The random effect model is used if one of the above assumptions of fixed effect model is not achieved (Gujarati, 2003).

The block effect in random effect model can be represented as:

$$\beta_{0(i)} = \mu + v_i \qquad (i = 1,2,\dots,N)$$

Putting the equation (7) in equation (3) we get the random effect model below:

$$Y_{it} = \mu + \sum_{j=1}^{k} \beta_j X_{j(it)} + v_i + \epsilon_{it} \qquad (i = 1,2,\dots,N \;\; t = 1,2,\dots,t) \qquad (8)$$

v_i Represent the cross sectional specific error component.

ϵ_{it} Represent the combined time series and cross section error component.

In this case $\epsilon_{it} \sim IID\left(0, {\sigma_{\epsilon_{it}}}^2\right), v_i \sim IID\left(0, {\sigma_v}^2\right)$ and the(ϵ_{it})is independent of the (v_i). In addition, the (X_{it})is independent of the (ϵ_{it})and (v_i)for all (i) and (t).

Let we have composite error term:

$$w_{it} = v_i + \epsilon_{it} \qquad (9)$$

Where $E(w_{it}) = 0$

$$\text{var}(w_{it}) = \sigma^2{}_v + \sigma^2{}_\epsilon$$

For estimating the parameters in the random effect model we use generalized least square estimation, we can not use (OLS) estimation because the estimations will be inconsistent.

For testing that which model fixed effect or random effect model is appropriate model we can use the Hausman test (William H. Greene-2003).

Hausman test:

In panel data analysis the hausman test can helps to choose between fixed effect model or random effect model the null hypothesis is that the preferred model is random effect the alternative hypothesis is that the preferred model is fixed effect model.

H_0: Random effect model is appropriate.

H_1: Fixed effect model is appropriate.

Where $$\lambda = 1 - \left(\frac{\sigma^2{}_v}{\sigma^2{}_v + T\sigma^2{}_\epsilon}\right)^{1/2} \tag{12}$$

Hausman statistic is calculated from the formula:

$$H = (\beta^{RE} - \beta^{FE})' \left[Var\left(\hat{\beta}^{RE}\right) - Var(\hat{\beta}^{FE}\right]^{-1} (\hat{\beta}^{RE} - \hat{\beta}^{FE}) \tag{13}$$

Where (β^{RE}) and (β^{FE}) are the vectors of coefficient estimates for the random and fixed effects model respectively.

This statistic is $(x^2{}_k)$ distributed under the null hypothesis. The degrees of freedom (k) equal the number of factors.

The statistic, computed above is compared with the critical values for the (x^2)distribution for (k) degrees of freedom. The null hypothesis is rejected if the Hausman statistic is greater than its critical value.

The random effect model is a model between the pooled regression model and the fixed effect model notes that in below equation:

$$y_{it} - \lambda\bar{y}_i = \beta_0(1 - \lambda) + \beta_1(x_{it} - \lambda\bar{x}_i) + v_{it} - \lambda\bar{v}_i \tag{14}$$

From equation (13) when ($\lambda = 0$) the model goes back to pooled model and when (λ=1) the model becomes a fixed effect model.

- The (λ) is always between (0) and (1) in the random effect model.
- As ($t\to \infty$) the fixed effect and random effect model are equivalent since ($\lambda\to 1$).

Chapter Three

Application

3-1 Data description:

Before we proceed the analysis it may noted that the panel data in this study is balanced panel data because the number of time observations is the same for each individual the data here also called long panel data in long panel the number of cross section or individual unit is smaller than the number of time period. Suppose we want to estimate model of wheat production giving in relation to the variables area of agriculture and rainfall.

The data of this study is yearly wheat production and area that used with yearly rain fall in the cities of sulaimani, Hawler and Duhok during the years (1992 - 2011), it is contain (60) observation and they are measured by tons, acres and milliliter, these data are time series data and cross section data which means that we have more than one dimension which is called panel data, was taken from the ministry of agriculture and metrological office of weather in sulaimani, Hawler and Duhok. The response variable is wheat production and the explanatory variables are area that used for producing wheat and the rain fall amount.

The data that had been collected as described in the previous are used to perform a Panel data analysis by using Eviews9 software, after standardizing all variables under study and the results are presented in these steps below:

First step: Estimating the pooled regression or pooled OLS model as initial model, its results are shown in table (3-1) below.

$$\hat{y}_{it} = -0.229236+0.380433X_1+0.273459X_2 \quad (1)$$

Table (3-1): The summary result for pooled OLS model.

Table (3-1)

Shows the summary result for pooled OLS model

Model	Coefficient		P-value	Mse	Akaike	R^2	Adj R^2	DW
Pooled OLS	b_0	-0.229236	0.0001	8.9	1.03	0.48	0.46	1.83
	b_1	0.380433	0.0000					
	b_2	0.273459	0.0000					

From above table the column of the coefficient represents the estimated parameters of area per acres and rainfall per millimeter which are equals to (0.380433 and 0.273459) respectively that means the yields of wheat is increased by (0.380433) per acres of area and it is increased by (0.273459) per mm of rainfall as it is obvious from the P-value of the estimators is less than (0.05) it implies that both estimators are statistically significant, in another word the null hypothesis that says $(\beta_1 = \beta_0 = 0)$ is rejected. The determination of coefficient (R^2) is (%48) that means both explanatory variables capable of explaining %48 of the response variable but the other %52 is related to unknown information that the researcher couldn't get them. The comparison measurements (Mse and AIC) are equal to (8.9 and 1.03) respectively they should be used as a comparison measures among the candidate model to get the best postulated model that fit the data under study.

The hypothesis of autocorrelation problem have been tested depending on Durbin-watson test which is clear from the above table the calculated D.W test value is (1.83) and it is greater than the $(d_L = 1.351)$ from D.W table it implies that the model doesn't suffers from A.C problem .

The predicting values after running pooled model are shown in the graph in figure (1).

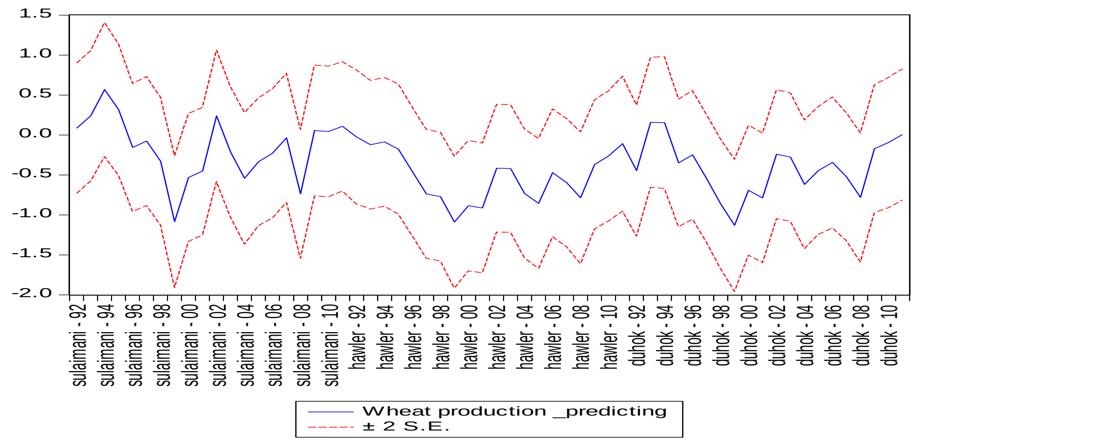

Figure (1) explain predicting of wheat production using pooled OLS model

To sum up figure it is clear that the predict values are fall between the upper bound and lower bound.

Second step: Estimating the Fixed effect model as initial model, its results are shown in table (3-2) below.

$$\hat{y}_{it} = -0.204358+0.417033X_1+0.370762X_2 \qquad (2)$$

Table (3-2): The summary result for fixed effect model.

Table (3-2)

Shows the summary result for fixed effect model

Model	Coefficient		P-value	Mse	Akaike	R^2	Adj R^2	DW
Fixed effect model	b_0 b_1 b_2	-0.204358 0.417033 0.370762	0.0013 0.0029 0.0077	2.9	0.63	0.82	0.72	1.88

From above table the column of the coefficient represents the estimated parameters of area per acres and rainfall per millimeter which are equals to (0.417033 and 0.370762) respectively that means the yields of wheat is increased by (0.417033) per acres of area and is increased by (0.370762) per millimeter of rainfall as it is obvious from the P-value of the estimators is less than (0.05) it implies that both estimators are statistically significant, in another word the null hypothesis that says ($\beta_1 = \beta_0 = 0$) is rejected. The determination of coefficient (R^2) is (%82) that means both explanatory's variable capable of explaining %82 of the response variable but the other %18 is related to unknown information that the researcher couldn't get them. The comparison measurements (Mse and AIC) are equal to (2.9 and 0.63) respectively.

The hypothesis of autocorrelation problem have been tested depending on Durbin-watson test which is clear from above table the calculated D.W test value is (1.88) and it is greater than the ($d_L = 1.351$) from D.W table it implies that the model doesn't suffers from A.C problem .

The predicting values after running fixed effect model are shown in the graph in figure (2).

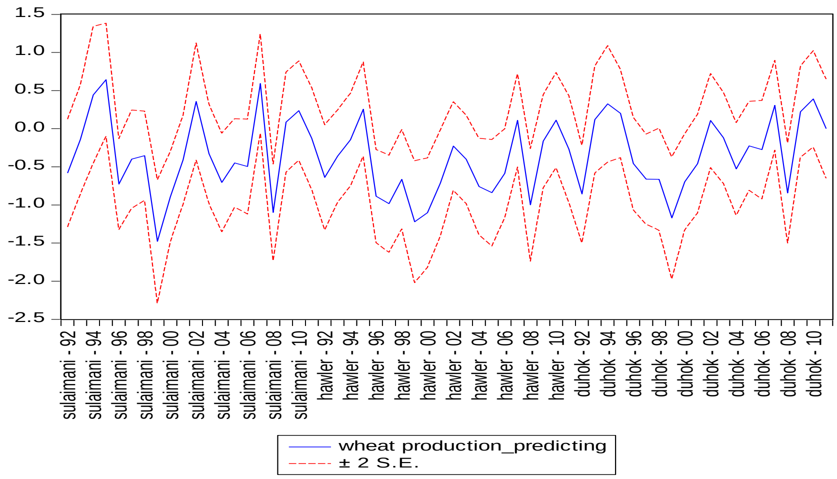

Figure (2) explain predicting of wheat production using fixed effet model

To sum up figure it is clear that the predict values are fall between the upper bound and lower bound.

Third step: Estimating the random effect model cross section random and period fix as initial model, its results are shown in table (3-3) below.

$$\hat{y}_{it} = -0.229236+0.380433X_1+0.273459X_2 \qquad (3)$$

Table (3-3): The summary result for random effect model cross section random and period fix.

Table (3-3)

Shows the summary results for random effect model

Model	Coefficient		P-value	Mse	Akaike	R^2	Adj R^2	DW
Random effect Model	b_0	-0.229236	0.0001	8.9	1.03	0.48	0.46	1.83
	b_1	0.380433	0.0000					
	b_2	0.273459	0.0000					

From above table the column of the coefficient represents the estimated parameters of area per acres and rainfall per millimeter which are equals to (0.380433 and 0.273459) respectively that means the yields of wheat is increased by (0.380433) per acres of area and is increased by (0.273459) per millimeter of rainfall as it is obvious from the P-value of the estimators is less than (0.05) it implies that both estimators are statistically significant, in another word the null hypothesis that says ($\beta_1 = \beta_0 = 0$) is rejected. The determination of coefficient (R^2) is (%48) that means both explanatory's variable capable of explaining %48 of the response variable but the other %52 is related to unknown information that the researcher couldn't get them. The comparison measurements (Mse and AIC) are equal to (8.9 and 1.03) respectively.

The hypothesis of autocorrelation problem have been tested depending on Durbin-watson test which is clear from above table the calculated D.W test value is (1.83) and it is greater than the ($d_L = 1.351$) from D.W table it implies that the model doesn't suffers from A.C problem .

The predicting values after running random effect model are shown in the in figure (3).

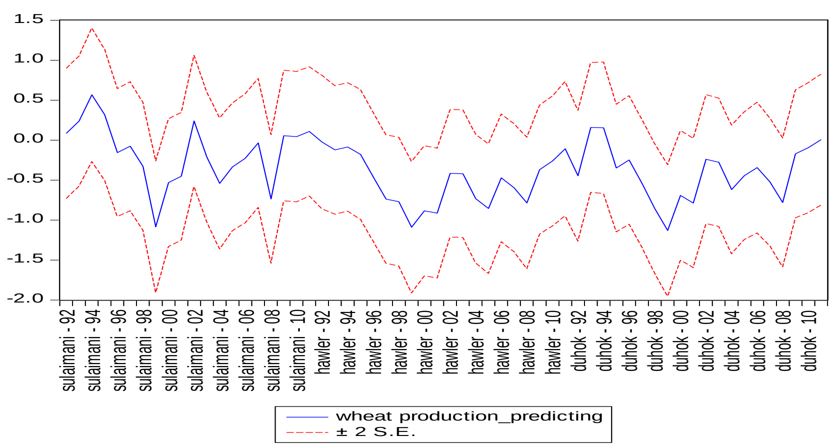

Figure (3) explain predicting of wheat production using random effect model

To sum up figure it is clear that the predict values are fall between the upper bound and lower bound.

Fourth step: Using Hausman test to test which model fixed effect model or random effect model is appropriate model.

H_0: random effect model is appropriate.

H_1: fixed effect model is appropriate.

The result of Hausman test is shown below it is output of Eviews9 software:

Correlated Random Effects - Hausman Test
Equation: Untitled
Test cross-section random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random	6.112543	2	0.0471

Cross-section random effects test comparisons:

Variable	Fixed	Random	Var(Diff.)	Prob.
X1_	0.425029	0.380433	0.000356	0.0182
X2_	0.289839	0.273459	0.001538	0.6762

As shown above we get a statistically significant p-value = 0.0471 which is less than 0.05 for that we reject the null hypothesis which says random effect model is appropriate model.

Chapter Four

Conclusions and Recommendations:

Conclusion:

After applying these three models pooled OLS, fixed effect model and random effect model on the data and consideration we conclude that:

1- In case we have panel data the pooled regression model can not be used because it does not detecting all possible pattern in the data.
2- The pooled model with panel data can not gives adequate results even the pooled model does not contain any econometrics problem.

3- The best panel model with a high performance that represents the behaviors of response variable and explanatory is the fixed model with (R^2=%82) and (Mse=2.9) and the minimum Akaike information (AIC=0.63).
4- From the figure (2) we can conclude that the prediction plot have the best presentation comparing with the other remains models see the fluctuations for the upper bound and the lower bound for the response prediction using fixed effect model (2).
5- From the Hausman test we conclude that the data is homogenous because the null hypothesis that says random effect model is appropriate has been rejected.

Recommendations:

We recommend the ministry of agriculture that:

1- We propose that the government through the Ministry of Agriculture to conduct a study in the provinces to determine the proportion of rainfall in each province for the necessity of wheat production.
2- Places that have a lot of agricultural land, but the rain is low the government should help them build the artesian well and give them the electricity required to have sufficient resources for agriculture.
3- The government should encourage farmers to grow agriculture by reducing production from outside the country.
4- We recommend the researchers to use the models of panel analysis with panel data to get adequate results.

References:

1- Alain Pirotte and Jean-Loup Madre, 2011. Car Traffic Elasticities:A Spatial Panel Data Analysis of French Regions, Journal of Transport Economics and Policy, Vol. 45, No. 3, pp.341-365.
2- Badi H. Baltagi, 2006. Panel data econometrics (Theoretical contribution and Empirical Application), book, Syracuse University, USA.
3- Badi H.Baltagi, 2005. Econometric Analysis of panel data, Third edition, British library cataloguing in publication data. Book.
4- Cheng Hsiao, 2007. Panel data analysis—advantages and challenges, Department of Economics, University of Southern California, Los Angeles, USA.

5- Desiderio Romero-Jordán, Pablo del Río and Mercedes Burguillo, 2014. Modelling Fuel Demand of Passenger Cars in Spain: A Dynamic Panel Data Analysis Using the Generalised Method of Moments, Journal of Transport Economics and Policy, Vol. 48, No. 2, pp. 315-332.

6- Edward W.Frees, 2004. Longitudinal and panel data (Analysis and applications in the social sciences), University of Wisconsin, Madison. Book.

7- Erik Biorn, 2017. Econometrics of panel data (Methods and Applications), book, Oxford University, United Kingdom. Book.

8- Ian McRae and James R. G. Butler, 2014. Supply and demand in physician markets: a panel data analysis of GP services in Australia, International Journal of Health Care Finance and Economics, Vol. 14, No. 3, pp. 269-287 Published by: Springer.

9- Jian guosun and Xing giu zhao, 2013. Statistical Analysis of panel count data, University of Missouri, Columbia and the hong Kong polytechnic University, Hong Kong. Book.

10- Lahcen El Kharti, 2014. The determinants of financial performance of microfinance institutions in Morocco: a panel data analysis, Savings and Development, Vol. 38, No. 1, pp. 27-44.

11- M. Zakir Saadullah Khan and M. Ismail Hossain, 2012. Determinants of Trade Balance of Bangladesh: A Dynamic Panel Data Analysis, the Bangladesh Development Studies, Vol. 35, No. 2, pp. 45-65.

12- Stéphane Bonhomme and Elena Manresa, 2015. Grouped patterns of heterogeneity in panel data, Econometrica, Vol. 83, No. 3, pp. 1147-1184 Published by: The Econometric Society.

13- William H. Greene, 2003. Econometric analysis, New York University, book.