ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc. THESIS

Darya Latif RASUL

PANEL DATA ANALYSIS AND MODELLING IN THE FIELD OF AGRICULTURE

DEPARTMENT OF STATISTICS

ADANA-2021

This dissertation is dedicated to my sacrificial mother

ABSTRACT

MSc THESIS

PANEL DATA ANALYSIS AND MODELLING IN THE FIELD OF AGRICULTURE

Darya Latif RASUL

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCE
DEPARTMENT OF STATISTICS

Supervisor : Prof. Dr. Güzin YÜKSEL
Year: 2021, Pages: 75
Juries : Prof. Dr. Güzin YÜKSEL
: Prof. Dr. Gülsen KIRAL
: Prof. Dr. Hamza EROL

It is primary clear that data is the most important element of statistical analysis, for the same reason we focus on the data complementary. Thus, with the diversity of data the statistical results are more useful and illustrative.

Here our study aims to refer about panel data analysis and modelling which is based on panel data. Panel data also known as a longitudinal data in statistics and econometrics is that composed of observations of different cross section data's across the time span. In this study regression is run over two dimensional (cross section and longitudinal) panel data and with the test specifications for determining either to use and prefer pooled, fixed, random or mixed effects models.

After condensing the estimation methods, briefly it is concluded a discussion of results. Panel data analysis and modeling provides a complete and rich study for the purpose of our research area.

Key words: Panel data, fixed effect, random effect, mixed effect, wheat product

ÖZ

YÜKSEK LİSANS TEZİ

TARIM ALANINDA PANEL VERİ ANALİZİ VE MODELLENMESİ

Darya Latif RASUL

ÇUKUROVA UNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Prof. Dr. Güzin YÜKSEL
Year: 2021, Sayfa: 75
Jüri : Prof. Dr. Güzin YÜKSEL
: Prof. Dr. Gülsen KIRAL
: Prof. Dr. Hamza EROL

İstatistiksel analizin en önemli unsurunun veri olduğu açıktır, bu nedenle tamamıyla veri üzerine yoğunlaşırız. Dolayısıyla, verilerin çeşitliliği sayesinde istatistiksel sonuçlar daha yararlı ve açıklayıcı olurlar.

Burada çalışmamızda, panel veriye dayalı panel veri analizi ve modellemesi konularının incelenmesi amaçlanmıştır. İstatistik ve ekonometride boylamsal veriler olarak da bilinen panel veriler, bir zaman aralığı boyunca farklı yatay kesit verilerine ait gözlemlerden oluşan verilerdir. Bu çalışmada, iki boyutlu (yatay kesit ve boylamsal) panel verilere, havuzlanmış, sabit, rastgele veya karma etkili modellerin kullanımının tercihinde tanımlanan testler ile regresyon analizi uygulanır.

Tahmin yöntemleri üzerinde durulduktan sonra, kısaca sonuçların tartışılması ile noktalanır. Panel veri analizi ve modellenmesi, çalışmamızın amacı için eksiksiz ve zengin çalışma sağlamaktadır

Anahtar kelimeler: Panel verileri, sabit etki, rastgele etki, karma etki, buğday ürünü.

EXTENDED ABSTRACT

This study examines the power of panel data analysis as we use it for fitting panel models associated with production activities in agricultural field, mainly wheat production in countries of Turkey,Iraq and Iran. Panel data analysis is one of the most frequently used statistical techniques for data analysis and it is used in a wide range espicially in econometrics for analysing two dimentionals typically time and cross sectionals.

The data are usually collected over more than two dimensions (time, location, individuals), after data collection over time and same individuals, it is followed by running regression over these dimentions. We must mention that the error term (E_{it}) has got a big role in the analysis and decisions where assumptions of the error term guide us whether we should determine fixed effects, random effects or mixed effects model. Also we ought to mention that agriculture is measured as one of the most important part of economic sectors also it has effect on the lifehood sectors. As developments of agriculture field do not have only effect on the insurance of local foods itself, but it also has very significant effect on reducing jobelessness and improving many other monetaries and vital aspects in the country.

In this research, the annual wheat production amounts, cultivation areas and rainfall amounts data are used to make the best model for wheat product comparison by using statistical methods. These methods, named as panel data analysis and modeling, are widely utilized in social sciences and econometrics to analyze two dimensional typically cross sectionals and time. Mostly considerable benefits from usage of this method is to control heterogeniety. Here panel data analysis is suggested for which composed of cross section and time series, that may come across heterogeniety and causes in losing control. Besides it can cause a risk of obtaining biased results.

In order to compare the wheat product amount and get an idea about best circumstance which are decided by including the annual cultivated areas and rainfall amounts for each of the neighbour countries, Turkey, Iraq and Iran, here we use panel data analysis to fit panel data model for wheat production in these three countries.

Thus, for the best model specification on products, four approaches are used which are pooled regression model (PME), Fixed effects model (FEM), Random effects model (REM) and mixed effects model. Also for determination of the best model we saw that F restricted test and Hausman's test are necessary to be done.

As a concluding remark for our research, we can say that the best model with a high performance which represents the behavior of responses and explanatory variables, is the fixed effects model (fixed for cross section and period) with ($R^2 = \%96$) and (MSE = 0.0155).

GENİŞLETİLMİŞ ÖZET

Bu çalışma, tarım alanındaki üretim faaliyetlerinin, özellikle de Türkiye, Irak ve İran ülkelerindeki buğday üretimi için, panel modellere uygulanmasındaki panel veri analizinin gücünü inceler.

Panel veri analizi, veri analizi için en sık kullanılan tekniklerden biridir ve özellikle ekonometride genellikle zaman ve yatay kesit olmak üzere iki boyutlu analizler için yaygın olarak kullanılır. Veriler genellikle ikiden fazla boyutta (zaman, konum, bireyler) toplanır, zaman içinde ve aynı bireylerden veri toplandıktan bu boyutlar üzerinden regresyon uygulanır.

Hata teriminin (E_{it}), sabit etki, rastgele etki veya karma etki modelini belirlememiz gerektiğinde bize yol göstermede büyük rolünün olduğunu belirtmeliyiz. Ayrıca tarımın ekonomik sektörün en önemli parçalarından biri olarak ölçüldüğünü ve aynı zamanda yaşam sektörlerinde de etkisinin olduğunu belirtmek gerekir. Tarım alanındaki gelişmeler sadece yerel gıdaların sigortalanmasında etkili olmayıp, aynı zamanda ülkedeki işsizliğin azaltılması, diğer birçok mali destek ve hayati yönün iyileştirilmesi gibi konularda çok önemli etkilere sahiptir.

Bu araştırmada, buğday ürünü karşılaştırması için istatistiksel yöntemler ile en iyi modeli oluşturmada yıllık buğday üretim miktarları, ekim alanları ve yağış miktarları verileri kullanılmıştır. Panel veri analizi ve modellemesi olarak adlandırılan bu yöntemler, sosyal bilimlerde ve ekonometride genellikle zaman ve yatay kesit verisi olan iki boyutlu veride yaygın kullanılmaktadır.

Bu yöntemin kullanımının en önemli faydası, heterojenliği kontrol etmektir. Burada kesit ve zaman serilerinden oluşan, heterojenlikle karşılaşabilen ve kontrolü kaybetmeye neden olan panel veri analizi önerilmektedir. Ayrıca yanlı sonuçlar elde etme riskine neden olabilir. Burada buğday ürün miktarını karşılaştırmak ve komşu ülkeler, Türkiye, Irak ve İran'da, yıllık ekili alanlar ve yağış miktarları dahil edilerek kararlaştırılan en iyi durum hakkında fikir edinmede,

bu üç komşu ülkenin buğday üretimi için panel veri modeline uyacak panel veri analizini kullanıyoruz.

Bu nedenle, ürünlerle ilgili en iyi modelin belirlenmesi için, havuzlanmış regresyon modeli (PME), sabit etki modeli (FEM), rastgele etki modeli (REM) ve karma model olmak üzere dört yaklaşım kullanılır.

Ayrıca en iyi modelin belirlenmesi için kısıtlı F testinin ve Hausman testinin uygulanması gerektiğini gördük.

Araştırmamız için sonuç olarak, yanıtların ve açıklayıcı değişkenlerin davranışını temsil eden yüksek performanslı en iyi modelin R^2 = % 96 ve MSE = 0.0155 ile sabit etki modeli olduğunu söyleyebiliriz.

AKNOWLEDGEMENTS

First thanks to my god, my guide and by my side in every difficulty and crowd that have been on my way during my study (I love you my lord).

I am grateful to my supervisor. Prof.Dr. Güzin YÜKSEL for her support, devoted guidance, encouragements and moral boosting from the first day of my study till the end.

Special thanks to my mother, sisters and friends for their support and tolerance during the entire MSc. Education.

Ending by, I would like to thank my husband for his motivation and support whenever needed.

CONTENS **PAGE**

ABSTRACT I
ÖZ II
EXTENDED ABSTRACT III
GENİŞLETİLMİŞ ÖZET V
AKNOWLEDGEMENTS VII
CONTENS VIII
LIST OF TABLE XII
LIST OF FIGURE XIV
LIST OF ABBREVATIONS XVI
1. INTRODUCTION 1
1.1. Basic Concepts of the Study 1
1.2. Literature Survey on Panel Data Analysis and Modelling 4
1.3. The Study Intentions 7
1.4. Hypothesis 7
1.5. Layout of the Thesis 8
2. THEORETICAL DETAILS 9
2.1. Introduction and Explanations 9
2.2. Types of Panel Data and Variables 10
2.3. Advantages of Panel Data 12
2.4. Approaches to Linear Regression Analysis 13
2.5. Cross Section Data Analysis 14
2.6. Time Series Analysis 14
2.7. Assumptions of Panel Data Models 15
2.8. Panel Data Analysis 15
2.9. Panel Data Models 17
2.9.1. Pooled Model 17
2.9.2. Fixed Effects Model 18

2.9.2.1. Assumptions on Fixed Effects Model 21
2.9.2.2. Main Advantages of Fixed Effects Models 22
2.9.2.3. Restricted F-Test 22
2.9.3. Random Effects Model 22
2.9.3.1. Generalized Least Squares Method 24
2.9.3.2. The Variance Component Modules Estimators 25
2.9.3.3. Advantages and Disadvantages of Random Effects Model . 27
2.9.3.4. Hausman's Test 28
2.9.4. Mixed Effects Models 29
2.10. Auto-Correlation 32
2.11. Heteroscedasticity 33
3. APPLICATIONS AND ANALYSIS 35
3.1. Introduction 35
3.2. Data Description 35
3.3. Variables of the Study 36
3.3.1. Dependent Variable 36
3.3.2. Independent Variables 37
3.4. Application 38
3.4.1. Heteroscedasticity Test 38
3.4.2. Pooled Model 39
3.4.3. Fixed Effects Model Estimation (fixed cross section) 43
3.4.4. Fixed Effects Model Estimation (fixed cross section and period) 46
3.4.5. Random effects model estimation (random for cross section) 50
3.4.6. Mixed Effects Model Estimation (random cross section and fix period) 54
3.4.7. Mixed Effects Model Estimation (random period and fix cross section) 57
3.4.8. Models Fitting Results for the Panel Data 61
3.4.9. ANOVA Table 61

3.4.10. Comparison of the Models 63
3.4.10.1. Restricted F Test 63
3.4.10.2. Hausman's Test 63
4. CONCLUSION AND POSSIBLE RECOMMENDATIONS 65
4.1. Conclusions 65
4.2. Recommendations 66
REFERENCES 69
CURRICULUM VITAE 75

LIST OF TABLE **PAGE**

Table 3.1. Shows the summary of the result for the pooled model 39
Table 3.2. Shows the summary of the pooled model after removing AC problem..... 41
Table 3.3. Shows the summary result for fixed effect model fix for cross section..... 43
Table 3.4. The summary result for fixed effect model fix for cross section after removing AC..... 45
Table 3.5. The summary result for fixed effect model, fixed for cross section and period..... 47
Table 3.6. The summary result for fixed effect model fixed for cross section and period after removing AC..... 49
Table 3.7. The summary result for random effect model random for cross section..... 51
Table 3.8. The summary of random effect model after removing AC problem..... 53
Table 3.9. The summary result for mixed effects model, random for cross section and fix for period..... 54
Table 3.10. The summary of mixed effects model, random for cross section and fixed for period, after removing AC. 56
Table 3.11. The summary result for mixed effects model, period random and cross section fix. 57
Table 3.12. Shows the summary of mixed effects model, period random and cross fix, after removing AC problem. 59
Table 3.13. The models summary and comparisons 61
Table 3.14. The variance analysis of all the models..... 62
Table 3.15. The result of the Hausman's test. 64

LIST OF FIGURE **PAGE**

Figure 3.1. The graph of wheat production in the countries Turkey, Iraq and Iran. 36
Figure 3.2. Graph of the area for producing wheat for each of the countries Turkey, Iraq and Iran. 37
Figure 3.3. Graphs of rainfall for the countries Turkey, Iraq and Iran. 38
Figure 3.4. Graph of predicting wheat production for the pooled OLS model. 40
Figure 3.5. Graph of predicting wheat production for the pooled model after removing AC. 42
Figure 3.6. Graphs of predicting wheat production for fixed effect model fixed for cross section..... 44
Figure 3.7. Graph of predicting wheat production for fixed effect model fixed for cross section after removing AC problem. 46
Figure 3.8. Graph of predicting wheat production for fixed effect model fixed for cross section and period..... 48
Figure 3.9. Graph of predicting wheat production for fixed effect model, fixed for cross section and period, after removing AC problem. 50
Figure 3.10. Graph of predicting wheat production for random effect model, random for cross section..... 52
Figure 3.11. Graph of predicting wheat production for random effect model after removing AC problem 53
Figure 3.12. Graph of predicting wheat production for mixed effects model, random for cross section and fixed for period..... 55
Figure 3.13. Graph of predicting wheat production for mixed effects model, random for cross section and fixed for period, after removing AC.... 56
Figure 3.14. Graph of predicting wheat production for mixed effects model, period random and cross fix. 58

Figure 3.15. Graphs of predicting wheat production for mixed effects model, period random and cross fix, after removing AC. 60

LIST OF ABBREVATIONS

PLM	: Pooled Regression Model
FEEM	: Fixed Effect estimation Model
REEM	: Random Effect Estimation Model
OLS	: Ordinary Least Square
GLS	: General Least Square
LSDV	: Least Square Dummy variable
BLUE	: Best Linear Unbiased Estimator
LUE	: Linear Unbiased Estimator
RRSS	: Restricted Residual Sum of Squares
URSS	: Unrestricted Residual Sum of Squares
Ω	: Variance- Covariance Matric
$\otimes$	: Kronecker Product
BQU	: Best Quadratic Unbiased
DW	: Durbin Watson d test
d.f	: Degree of Freedom
R^2	: Determination of Coefficient
A.C	: Autocorrelation
AIC	: Akaike Information
SSE	: Sum Square Residual
Mse	: Mean Square Error
WLS	: Weighted Least Square

1. INTRODUCTION

1.1. Basic Concepts of the Study

Certainly, we know that agriculture is one of the most important parts of general economy of countries, as it plays a critical role in the entire transition of a given economy.

Agriculture can be measured as a backbone of economic system of any country.

In addition of providing food and raw materials, we must mention that agriculture also secure jobs and employment opportunities to a very large percentage of population.

Unfortunately, in developing countries derived oil industry has become one of the prime reasons that all the governmental wings have complete control all overwhelmed by the idea of pure greediness, beside that fertile agriculture laborers also can reinstate employment.

In the past, worldwide had more beneficial and healthier land for agriculture due to having log cabins for farming such as fertile land, glasshouse, stream, watering and overall a better weather situation, particularly before primal global warming wave.

Additionally, after that some of the empirical countries have tough human racing and their major role in agriculture and how to extend their income into building history and man-power construction.

Wheat is measured as the most important product of agriculture that has a big effect on economy and marketing, that’s why all the developed countries try to increase their wheat product and shop worldwide also trying to find its alternative.

Each of Turkey, Iraq and Iran countries have a leading source of wheat product as their climate facilities makes the way easier for the amount of product duplication.

Currently, the progression of agriculture in these countries are particularly prepared for marketing and with the increasing of population also internal industries all begin.

Taking into account, for each of the countries of Turkey, Iraq and Iran in 2019 the total wheat consumption is (20, 7.7, 16.9) thousand metric tons with the population of (83.2, 40.4, 82.9) millions respectively.

Today, wheat is produced in more land areas and more than any other commercial crops, which stays as the most important grain source for human being.

In the last twenty years, wheat has a powerful effect on the economy of these countries. Besides, many families could earn from the productions.

What more, in 2019 each of the countries Turkey, Iraq and Iran are planted wheat by the area of (68463, 6331, 61195) thousand hectares respectively.

After the development of agriculture, the statistical models have become more effectively used, for example multiple linear regression models are more precise than the experimental design models of agriculture.

For our study the panel data analysis is used to fit the panel data model for wheat production among Turkey, Iraq and Iran countries and making a statistical comparison of the wheat production between three locations to announce the decision maker able to determine the best location for the best results of the product and advisement to the other locations, in order to improve their wheat product level into a more sufficient level.

Accordingly, panel data has become widely available in both the developed and developing countries.

Panel data econometrics is one of the most accurate and interesting fields of investigation in today's econometrics, as it controls heterogeneity of individuals where biasedness is obtained due to lack of control this heterogeneity.

Maximum likelihood estimator is used for estimating these dimensional statements run under control to gain these models by each of random effects, fixed effects or mixed models in attendance of generalized least square estimator, which

is a statistical method utilized in econometrics to analyze two dimensional of cross section data and time series data.

The data are collected over time and same individuals which are explained by regression.

Definitely panel data analysis processed for examining particular subjects in so many wide platforms, so often maintained over a definite time frame.

Briefly, in econometrics panel data analysis is used to study the behavior of trade and income. In social sciences, panel data analysis is approvable for lengthwise analysis.

In political science, it is used to study political activities of organizations and factions over time. Also in health studies to examine certain features generally followed over time, same for psychology and sociology.

However, in educational studies around students, graduates or employed over time.

When repetition occurs in individuals of certain cross sections, panel data analysis allows the researcher to act with rapid changes with short time series.

Adaption of cross section with time series can expand the quantity and quality of data in such a manner cannot be obtained by using one of these two dimensions.

In addition, panel data analysis provides regression analysis with two dimensional and temporal lengths, the spatial dimension affects the set of cross sectional and time dimensions.

Study on the regression models to fit on econometric models is more complicated than simple cross sectional data set, as they are used in applies works progressively, ever more Panel data analysis is a very restorative circumstance in econometrics today.

1.2. Literature Survey on Panel Data Analysis and Modelling

In literature many studies have been made that deal with panel data analysis and modelling and making the application of some special extrapolation process on it. In (1986) Barbara E Kahn, Manohar U Kalwani and Donald G Morrison offered a framework that was composed of seven simple models which were tested on panel data, composed of sixteen brands in five different categories. Besides, the running test to consider about the effect of violations of the model assumptions.

In (1992) Fumio Hayashi has presented an article by Keane and Runkle which included three contributions to the econometrics of panel data. Step one suggests a new estimator under prearrangement. at the second step suggests a Hausman specification test of erogeneity of instruments and generalizes the standard Hausman test in the absence of the fixed effect. And the third step it contributed an experiment in which the standard Hausman test can be highly deceitful as instruments are not exactly determined.

In (1997) X Liu, H Song, Y wei, P Romilly have had a study about country characteristics and foreign direct investments in china which alone deals with error component model that can be used to analyze in any political, economic and cultural determinants, the panel data takes the period between (1983 and 1994) with 17 regions. The result showed that bilateral trade, cultural regional difference make change in market size, wage rate and exchange rate that are necessary determinants.

In (2004) L. M. Bhole and Jitendra Mahakud estimated both time series and panel data models for empirically classifying the elements of trade credit of the entire economy, public limited companies, private limited companies and foreign companies in India from 1966-67 to 2000-01. The study finds that the government sector has continued a considerable user of trade credit through the entire period. The nature of behaviors of trade credit and the changes in it over the years clearly depend upon the type of ownership of the companies.

Also, in (2008) Julie A. Phillips and David F. Greenberg fixed the idea about the special cases of a more general model on each of fixed and random effects models for pooled cross-sectional and time series data, and latent growth curve models for panel data. Also they separated longitudinal data into three main types. Panel data (cross-sectional, time series data) contain observations on multiple cases over multiple time periods while Time-series. Data consists of various observations over time for a single case and several statistical techniques are available for analyzing panel data and the use of fixed or random effects for cases.

Elsewhere, in (2010) Keuntae Kim and Joel E. Cohen managed panel data analysis techniques to realize causes of international immigration in (17) wealthy nations and international emigration from (13) of those (17) wealthy nations between (1950 and 2007) in (77,658) remarks from multiple sources. The study examined panel data analysis to under come heteroscedasticity and autocorrelation in those panels.

In (2011) Ivan A. Canay contribute some conditions which determine a quintile regression model with fixed effects and offers a simple transformation of the data to remove the fixed effects by a theory, that these special effects are location shifters, in which the new estimator is consistent and nearly normal as both η and T grow.

More often, Alain Pirotte and Jean-Loup Madre in (2011), examined panel data set of twenty-one French cities which used the resistance of car traffics. The study clarifies that the useful way of using panel data to control heterogeneity where time series and cross-sections examine and do not control such heterogeneity that run the risk of obtaining biased results. Conditions of random or fixed effects estimated using the maximum likelihood estimator. This also presents the estimates of feasible Generalized Least Squares estimator and performed a Hausman's test to make the fixed and random effects specification of the spatial models.

M. Zakir Saadullah Khan and M. Ismail Hossain in (2012) covered dynamic panel data analysis techniques. It was covering Bangladesh's employment with its 50 major trading partners for over 26 years using Unrestricted Error Correction Mechanism (UECM) and Generalized Method of Moments (GMM) estimator. The usage of such technique was related to the results of the unrestricted error correction model. In (2014) All Zakir Husain, Mousumi Dutta and Nidhi Chowdhary have managed panel data on 216 countries for the time interval from (1980 to 2009). By means of dynamic panel data models also instrumental variable models for panel data and the random effects model (REM) was the significant idea. By the same procedure of Hausman test on the panel data and basing on the difference between the fixed effects and random effects estimators and the estimated variance-covariance matrix of variations.

Also Muhammad Baraka and Muhammad Jabury in (2014) used panel data to manage method of generalized moment's valuation to study on the impact of exchanging rate commands on inflation of 50 countries, over the period from (1980 to 2008). analysis results show rising in regimes that negatively and significantly have been targeted.

In (2016) Noura muhammady and Muhammad Al-Saghir Quraishy have studied about panel regression model to realize any changes on the independent variable measured by the properties of the board on the dependent variable, that is the financial performance. They realized a necessary positive influence of the independence of the board on their work performance.

Lastly but not least; in April (2020) Nasrin Salehnia,Nargis Salehnia, Ahmad Saradari and Sohrab kolsoumi in their study about rain fed wheat yield, the prediction with approach to pooled modeling panel data, that draught indicators through the time period between (1990-2016). The study approved that the pooled model is a real implementer for wheat yield prediction.

1.3. The Study Intentions

In our study we are intended to investigate the following points;

1- Analyzing panel data to fit panel data models for wheat production in the three countries of (Turkey, Iraq and Iran).
2- Deriving Pooled regression model (PRM), Fixed effect estimation model (FEEM), Random effect estimation model (REEM) and Mixed effects model.
3- Using F-restricted test and Hausman's test for testing these four types of models and selection of best model acting as the best relationship of panel data set.
4- Choosing ideal model for detecting and organizing the variable that has incidental effects on wheat production.
5- Selecting out most suitable location for wheat production with regardless of any other outstanding factors that can help us in the recommendations for the other locations.

1.4. Hypothesis

The hypothesis of the study including;

1- Time interval of wheat production shows a great rate of increasing in the measurement of wheat production level in all three countries of (Turkey, Iraq and Iran).
2- Each location separately targets truly award ability to inform their need for domesticating wheat product.
3- Cross sections with locations (Turkey, Iraq and Iran) have optimal changes in wheat product which are variety in their product.
4- Detection of any relationship between individual's distinct changes in the regresses or factors and any changes in the dependent variable the estimated

models can be practiced, same for any unusual changes of dependent variable under the effect of locations and times together.

1.5. Layout of the Thesis

In our thesis we have four main chapters;

Chapter one consists of general introduction and explanation by giving the main concepts of the study, Literature survey on panel data analysis and modeling also the study intentions and hypotheses. The following chapters are organized as follows:

Chapter two consists of theoretical details of the parts that are brief explanations about approaches to linear regression analysis and panel data analysis and modelling with required tests. In third chapter we have the data description and applications that includes a brief study on tests and analysis that are required for obtaining optimal results. Finally, the fourth chapter is a reaping of our work by giving the study conclusions with the suggestions and some possible recommendations.

2. THEORETICAL DETAILS

2.1. Introduction and Explanations

Statistics is known as the main circumstance for data collection, that associated with abstraction and analysis of data to express about any item related to human's daily life.

Econometrics relates the quantitative application of statistical and mathematical models benefiting from collected data to promote theories, hypothesis testing and trends forecasting in economics.

In econometrics and statistics term of panel data, longitudinal data or combined time series/cross section data are used to express data sets that contains repeated observations on a liberty of variables from a set of observation entity.

These observations envelop the temporal and the spatial dimensions concurrently, that is consistent of two types of data which are micro data and macro data sheets.

Mainly panel data is that data with categorical and organizing structure.

Panel data econometric is measured as most interesting and important field of inquiry in today econometrics with an ancient history returns to the year of 1950.

As the combination of cross section data and time series data is called panel data that can be collected on the same unit for two or more time periods, like collecting data on the same (individuals, jobs, schools, hospitals, cities, countries, firms and etc.).

The most powerful reason for using panel data is providing a large number of data points those result in increasing the degree of freedom and absence or decreasing in collinearity in explanatory variables. So on causes in the improvement of estimates efficiency.

More exciting reason, panel data makes facilities for the researchers or analysts to study on more serious and important economical questions that are

probably impossible to focus on the aim of the study by only cross section or time series data set. So the researchers and analysts can get a better parameter estimates, results in more precise and unbiased estimates, because more data means more variation and more information can be performed.

If the number of cross section units is (N) and the number of time periods is (T), then we have (N*T) observations, achieving two types of data means two types of variations are rolling in the study and enables the researchers to get more informative results.

Here variations accumulated in three major classes, which are ;

- Overall variation that happens over both time and individuals.
- Within variation is a variation within individuals but over time, causes changes in a unit over time periods, these results in proper estimates as it may be used to control the unobserved variables which only vary among units and not over time period.
- Between variations occur between individuals that lead to differences between units for the same time period.

Here the estimates are more likely to be unbiased because the within variation may control for unobserved variables that vary across units but not in time period, so on distracting any unobserved variable leads for permitting biased results.

2.2. Types of Panel Data and Variables

- The most necessary and reliable type of panel data is balanced panel data that in all time periods the same individuals are detected, more easily briefing in that we can say that a data set that includes all observed elements in all time frames.

Consequently, if a balanced panel data contains (N) members and (T) time periods, the number of observations (n) is (N*T), so a balanced panel set has a matrix form also it can be represented as three dimensional matrices because three subscripts may be ordered for each of the individuals those are number of units, time periods and for variable number.

Usually N is larger than T, but for geographic unit's panel, it is perhaps opposite. But when speech comes to variables, we notice that not all variables are vary both across observations over a time frame, for example for individuals (birth year, education period, gender), such variables are presented by individuals specific. And there may be variables which don't vary across individuals, for example (interest rate, tax parameter), these are represented as time specific or period specific.

The important term for us, when variables are seem to be variables across both individuals and time periods which are called as two dimensional variables, examples for such two dimensional variable(income and consumptions, production and labor input for firms).

- The second extremely important type is unbalanced panel data, characterized by different units of observation in all time periods.

The reasons for unbalanced panel's occurrence:

1. Data entry and data extract from the basic data set.
2. Insensible absence of observation in the time period.
3. Endogenous specifications, when the model is unable to explain features as a result of partially determination of units by the variables.

All these reasons cause in complicated parameters estimation and explanations if the mechanism specification refused for modeling and design of inference method.

So any incorrect selection issue is potentially a serious problem for any panel dataset.

2.3. Advantages of Panel Data

There are many significant and useful reactions of panel data that we must mention about in the following points:

1- Panel data are more accurate and measurable data as compared to using cross section or time series data alone
2- Panel data includes more information and more variables than pure time series data or cross section data alone.
3- Panel data can estimate and model for common and special behaviors of allover the group of data at the same time.
4- Also panel data enables us to test more complex behavioral models than cross section data models or time series models alone.
5- Micro panel data collected under any circumstance for example (firms, households, employees, etc.) are more error free and exact than same variables on macro panels which results in biased or eliminated assumptions.
6- Panel data also may control heterogeneity of individuals in such a manner other pure data studies at a high risk of overlooked heterogeneity.
7- As panel data has a matrix form which is more variability so it is less collinear determination for the variables as compared to time series studies which are more detectable with multicollinearity.
8- Panel data with more degrees of freedom and more efficiency so more informative data which leads to produce a better parameter estimates.

9- Panels have more capability to study the dynamics of modification as compared to cross sectional data that are likely to forbid many changes, so income changes, salaries, jobs, employees, could be more efficiently studied with panel data analysis.

2.4. Approaches to Linear Regression Analysis

Linear regression is defined as the first arithmetic rule for understanding individuals for prediction and modeling in a way that the aspect of linear regression analysis is the exact form of model predictions.

Most of analysts think that it is the most accurate and important among any other types of analysis.

There are the exact assumptions of linear regression analysis as the following:

1- Linearity in all parameters of the model.
2- Non stochastic regressors.
3- Expected value of error term is zero.
4- Homoscedasticity must be achieved that's all variances are constant.
5- No evidence of autocorrelation problem.
6- No evidence of multicollinearity.
7- No biasedness specification.

As regression analysis estimates the relationship between the component variables, shows if those relationships are significant between dependent variable and independent variables, also indicates the effect of many as more independent variables on dependent variable. Also helps us to discuss about forbidden and special impacts of variables measured on different scales.

These useful points about linear regression analysis makes analysts to consider about exclusion and rating the best set of variables for the strategy of predictive models.

2.5. Cross Section Data Analysis

It is analysis of data at a single steady point of time; the data set may be used to inform observations of multiple variables at a singular point in time.

A simple and clear example on cross section data set is gross domestic product of any country.

Another example is when we want to compare the financial direction of two or more governments or companies at a specific fixed time point.

Cross section data is benefited in economics and other social sciences mainly cross sectional data analysis may be used to compare any statement between different locations or objectives at same time interval.

Econometric analysis of cross section data is generally supposed that the data is independently generated and the observations are independent mutually.

2.6. Time Series Analysis

Time series analysis involves methods for analyzing time series data to bring out important statistics and informing the best results by the aid of disparate time periods.

The objective of time series modeling is to correctly study the past observations of a time series to make appropriate model that explains the main structures of the series.

Such models may help to produce future values for the time series to use in any forecasting method. As a result of necessary significance of time forecasting, the analyst or researcher must be aware of fitting an acceptable model to the essential time point. A time series model may be linear or nonlinear according to the real value of the time series linearity of previous observations.

2.7. Assumptions of Panel Data Models

1st. Uncorrelated errors: It can be seen normally that in each unit the errors can be correlated over the time period(T) and with (N) clusters that represents units, since there may be correlation between errors and any of clusters.

So with this correlation the ordinary least square estimation acts upon standard errors to be biased and with unbiased parameter estimates.

Here cluster-robust standard error is likely needed to be achieved as we need to fit for clustering with full estimates of standard errors.

2nd. No correlation of error with explanatory variables: It is possible for any correlation of error term with any explanatory variable, but when error is correlated with explanatory variable the biased estimate is produced due to ordinary least square estimator, here instrumental variable estimation is likely to be applied.

2.8. Panel Data Analysis

As we mentioned before when cross section data and time series data are unified together, we get a set of data which is called panel data. Variables of panel data have double subscripts as;

$$y_{it} = \alpha + X_{it}\beta + u_{it} \qquad (2.1)$$

$$i = 1,2,3,\dots,N$$

$$t = 1,2,3,\dots,T$$

That (i)subscribt represents cross sections like (individuals, countries, firms, etc.) and (t) subscript represents time like (years, months, days, etc.).

Where;

(N): the number of cross sections.

(T): the number of time periods.

(y_{it}): the dependent variable of(it^{th}) observation.

(X_{it}): the independent variable of(it^{th}) observation on (k) explanatory variable.

(α): the intercept for (1×1) scalar that is constant.

(β): the constant for ($K \times 1$)vector that measures effect of x on y.

u_{it}: the error term.

Mostly one-way error component is used in panel data analysis as;

$$u_{it} = \mu_i + v_{it} \tag{2.2}$$

Where:

μ_i: indicates the individuals- specific variation of y.

v_{it} : indicates the stochastic regression disturbance.

Here (μ_i) is showing the unobservable individuals-specific effect which is at a fixed time and it reflects any influence of individuals which is not entered the regression analysis.

The regression disturbance (v_{it}) varies with individuals and time as it denotes any reminder disturbance within the regression. Mainly (y_{it}) will represent output for productivity function that takes advantage of data across time and (X_{it}) will represent inputs.

So the unobservable individuals -specific effect are resembled by (μ_i) as we can rewrite the equation (2.1) as follows;

$$y = \alpha\iota_{NT} + X\beta + u = Z\delta + u \tag{2.3}$$

Under specified dimensions (NT):

(y) : ($NT\times1$) matrix

(X) : ($NT\times K$)

$Z : [\iota_{NT}, X]$

$\delta : (\acute{\alpha}, \beta)$

(ι_{NT}) : Vector of either N or T dimensions

Also (2.2) becomes;

$$u = Z_{\mu}\mu + v \tag{2.4}$$

As $u = (u_{11}, \dots, u_{1T}, u_{21}, \dots, u_{2T}, \dots, u_{N1}, \dots, u_{NT})$

With observations in a fitted state, individuals under initial index is over individuals with the faster index is over time;

$Z_{\mu}=I_N \otimes \iota_T$.

Where;

I_N: is an identity matrix of dimension (N).

ι_T: is a vector of ones of dimension (T).

Z_{μ} : is a selector matrix of ones and zeros.

$\otimes$: Kronecker product

The kronecker product sometimes is called matrix direct product, is the organizing of two matrices of undetermined size which causes partitioned matrix.

2.9. Panel Data Models

Panel data models give detail of the individual's structures across both time and cross sections, in the procedure of estimating the regression model, four types are utilized to analyze panel data, they are; The pooled model, Fixed effects model, Random effects model and Mixed effects model.

2.9.1. Pooled Model

Pooled regression model also known as common effect model is the model with constant coefficients; the concept of such constant coefficients is referred to

intercepts and slopes. Pooled model can be measured as a simple form among other panel data models as we pool the entire data then we deal with ordinary least squares regression model.

The common effect model or pooled model equation is comparably like the equation for ordinary least square as follows:

$$y_{it} = \alpha + \beta' X_{it} + u_{it} \quad (2.5)$$

$$i = 1,2,3, \ldots, N$$

$$t = 1,2,3, \ldots, T$$

As (N) is number of cross sections and (T) is the number of time periods.

2.9.2. Fixed Effects Model

This model guesses that the various intercepts can fit for difference between cross sections, so the method for estimating fixed effect model determines the usage of dummy variable technique in order to detect the difference in intercepts resulted from difference in the data, this performance is usually called least square dummy variable technique (LSDV).

In the fixed effect model, the ordinary least square again is used but a model must be assumed with constant intercept for each individual with inefficient time period, consequently it is needed to consider more models for controlling these differences. Here (μ_i) must be estimated as a fixed parameter, (v_{it}) as an independent stochastic disturbance is independent and identically distributed by ($0,\sigma_v^2$) and (X_{it}) independent of it for every (i) and (t). With an arranged set of individuals (N) fixed effect model could be specified briefly, but the inference is conditional on the particular cross sections. Here we must get the estimates of (α, β and μ) by ordinary least squares on the following equation;

$$y = \alpha_{lNT} + X\beta + Z_\mu \mu + v = Z\delta + Z_\mu \mu + v \quad (2.6)$$

Where;

(Z): is (NT× (K +1))

(Z_μ): is (NT× N)

As we see, we have a consistency with the number of individuals (N). As (N) is large it means we have a large number of individuals' dummies and a large matrix would be converted to ordinary least squares with dimensions that is addition of both individuals and number of dummies. Here one can get the least square dummy variables (LSDV) of estimators that is also known as within group estimator as only we can obtain the variation within each group as a result of this estimation.

For estimating (α) and (β) parameters, we may multiply the model by (Q) as it is a matrix and can be performed by deviations from individual's means.

Also we can achieve ordinary least square with transformed model as following;

$$y = X\beta + v \quad (2.7)$$

$$Q * (y = X\beta + v) \quad (2.8)$$

$$Qy = QX\beta + Qv \quad (2.9)$$

This (Q) matrix shows up the individual's effect subsequently ($QZ_\mu = Q_{lNT} = 0$) since ($PZ_\mu = Z_\mu$),here the regression ($\tilde{y} = Qy$) with typical element ($y_{it} - \bar{y}_i$.) on ($\tilde{X} = QX$) with typical element ($X_{it,k} - \bar{X}_{i.,k}$) as in the place of any individual we use the regressors so we have (K*K) inversion in the place of (N +K)× (N +K).

The ordinary least square estimator is resulting in;

$$\tilde{B} = (\acute{X}\ Q\ X)^{-1}\ \acute{X}Qy \tag{2.10}$$

We can get the estimators variation in the same way of partitioned inverse for (2.6) equation:

$$\begin{aligned} \text{Var}(\tilde{\beta}) &= \sigma^2{}_v\ (X'QX)^{-1} \\ &= \sigma^2{}_v(\widetilde{X}'\tilde{X}\ )^{-1}.\tilde{\beta}) \end{aligned} \tag{2.11}$$

Furthermore, applying generalized least square on (2.9) produces ($\tilde{\beta}$).
For the simple regression:

$$y_{it}\ =\ \alpha + \beta x_{it}\ + \mu_{\ i} + v_{it} \tag{2.12}$$

Also averaging over time gives:

$$\bar{y}_{i.} =\ \alpha +\ \beta\ \bar{x}_{i.} +\ \mu_i + \bar{v}_{i.} \tag{2.13}$$

As a result, subtracting (2.13) from (2.12) gives:

$$y_{it} -\ \bar{y}_{i.} =\ \beta\ (\ x_{it} -\ \bar{x}_{i.}\ ) +\ (v_{it} -\ \bar{v}_{i.}) \tag{2.14}$$

And, averaging across all observations in (2.8) gives:

$$\bar{y}_{..} =\ \alpha +\ \beta\bar{x}_{..} +\ \bar{v}_{..} \tag{2.15}$$

Since we must identify (α) and (μ_i) and we know they are fixed constant, they cannot be estimated without additional restriction, therefore the restriction ($\sum_{i=1}^{N} \mu_i = 0$) is presented.

Also a perfect multicollinearity or dummy variable duplication may exist, again to avoid this issue this restriction ($\sum_{i=1}^{N} \mu_i = 0$) is used on the dummy variable coefficients.

when speech comes to least squares dummy variables or fixed effect least squares, we come across a loss in degree of freedom as we estimate for (N-1) additional parameters collinearity due to intensify many dummies, also fixed effect estimator is not estimating time invariant variables like gender, religion and education, but can be estimated by Q transformation, when deviation from mean transformation is justified, in another word the time invariant variables are hidden due to dummies in (2.6) that means the regression relations cannot be achieved ,so on if (2.6) is the reliable model then (LSDV) is the best linear unbiased estimator (BLUE) that is related to the standard classic distribution of zero mean and variance-covariance matrix ($\sigma_v{}^2\ I_{NT}$).

Another point is when the time period have unlimited interval the fixed effect estimator is steady, but if time period is fixed and we have unlimited cross sections then because of any rising in number of individuals, arises the number of parameters the fixed effect of cross sections are not consistent but still the fixed effect of the parameters is steady. This issue is discussed by Neyman and Scott (1948) that is called incidental parameter problem.

2.9.2.1 Assumptions on Fixed Effects Model

One must focus on some assumptions that need to achieve in order to maintain unbiased estimates for ordinary least squares especially in large samples:

- Uncorrelated errors with entire observations of independent variable over time, otherwise if this assumption is not achieved, omitted variable bias will be encountered
- Allowing the individuals specific intercept to be correlated with the regressors (x_{it}).
- Independent and identically distributed variables across individuals.

2.9.2.2. Main Advantages of Fixed Effects Models

From previously mentioned assumptions clearly we can say that the main advantages of fixed effect estimation are for controlling unobserved confusion of variables which vary across units but not over time and eliminating correlation of errors that are resulted from the error component vi.

2.9.2.3. Restricted F-Test

By using F- test we can test for fixed effects that is testing for individual's effects and significance of dummies simply with ordinary least squares on the pooled model by restricted residuals sums of squares (RRSS) and unrestricted residual sums of squares (URSS) of the least square dummy variable.

$$F_0 = \frac{(RRSS-URSS)/(N-1)}{URSS/(NT-N-K)} \sim F_{N-1,N(T-1)-K} \quad (2.16)$$

Under the hypothesis:

H_0: pooled model is appropriate.

H_1: fixed effect model is appropriate

2.9.3. Random Effects Model

In random effects models, individual specific effects are arrested by a composite error term ($\alpha_i + u_{it}$) which assumes that individual intercepts extended

from a random distribution of possible intercepts. The random component of the error term α_i captures the individual specific effects in a different way from fixed effects models.

As one deal with a large population and drawing individuals would be from that population, then the random effects model could be guaranteed model specification method. Previously we discussed about pooled and fixed effects model and mentioned that when we have a large number of individuals the fixed effects models causes a huge loss in degree of freedom, so on we assume (μ_i) to be random, independent and identically distributed also independent of the (v_{it}) since (v_{it}) again independent and identically distributed. Another important point for this case is the estimator may be consistent only if the average of repressors over time are independent of errors and $((X_{it})$ in all time periods are independent of the (μ_i) and (v_{it}), so on consistent inference to the population that the sample is drawn from randomly, in fact this implies a homogenous variance as:

$$var(u_{it}) = {\sigma_\mu}^2 + {\sigma_v}^2 \tag{2.17}$$

For all (i) and (t)

We calculate the variance-covariance matrix as follows;

$$\Omega = \mathrm{E}(\mathrm{u\acute{u}}) = {\sigma^2}_\mu(I_N \otimes J_T) + {\sigma^2}_v(I_N \otimes I_T) \tag{2.18}$$

Another opinion ; Peter Kennedy in (A Guide to Econometrics,2008) discusses that the most information used by random effects makes them more efficient estimators, but correlation between 'x' and the error term causes bias as random effects assumes that α_i is uncorrelated with (orthogonal to) regressors. Also Angrist and Pischke in (Mostly Harmless Econometrics 2009) explain that they prefer fixed effect because the gains in efficiency are likely to be modest while the

finite sample properties of random effects may be worse. So on an important assumption for identification in fixed effects is that the most important sources of variation are time invariant (because information from time varying regressors gets differenced out).

2.9.3.1. Generalized Least Squares Method

We must calculate $(E(uú))^{-1})$ in order to obtain the general least square estimator of the regression coefficients. This is a large matrix for typical panels and is of dimension (NT× NT).

By derivation to get (Ω^{-1}) and $(\Omega^{-1/2})$:

$$\Omega = T\,\sigma_\mu{}^2\left(I_N \otimes J_T^{'}\right) + \sigma_v{}^2(I_N \otimes E_T) + \sigma_v{}^2(I_N \otimes J_T^{'})$$

Additive of same matrices gives the spectral decomposition representation of (**Ω**):

$$\Omega = \sigma_1{}^2 P + \sigma_v{}^2 Q \qquad (2.19)$$

Here $(\sigma_1{}^2 = T\sigma_\mu{}^2 + \sigma_v{}^2)$

Where $\sigma_1{}^2$: is the first unique characteristic root of (Ω) multiplying (N).

$\sigma_v{}^2$: is the second unique characteristic root of (Ω) multiplying $N(T-1)$.

Using the properties of (P) and (Q) verifying that:

$$\Omega^{-1} = \frac{1}{\sigma_1{}^2} P + \frac{1}{\sigma_v{}^2} Q \qquad (2.20)$$

$$\Omega^{-1/2} = \frac{1}{\sigma_1} P + \frac{1}{\sigma_v} Q \qquad (2.21)$$

Finally:

$$\Omega^r = (\sigma_1{}^2)^r P + (\sigma_v{}^2)^r Q$$

Obtaining (GLS) as weighted least squares by pre-multiplying the regression equation given in (2.3) by $\sigma_v\,\Omega^{-1/2} = Q + \left(\frac{\sigma_v}{\sigma_1}\right) P$ and performing (OLS) on the resulting transformed regression.

In this case, $(y^* = \sigma_v\,\Omega^{-\frac{1}{2}} y)$ has a typical element$(y_{it} - \theta y_{i.}^{'})$.

Where $(\theta = 1 - (\sigma_v/\sigma_1))$.

This transformed regression inverts a matrix of dimension (K +1) .

The best quadratic unbiased (BQU) estimators of the variance components arise naturally from the spectral decomposition of(Ω).

$$\text{And} \quad \hat{\sigma}^2{}_1 = \frac{\acute{u}\,P\,u}{tr(P)} = T\sum_{i=1}^{N} \bar{u}_{i.}{}^2 \,/\, N \tag{2.22}$$

$$\hat{\sigma}^2{}_v = \frac{\acute{u}\,Qu}{tr(Q)} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}-\bar{u}_{i.})^2}{N(T-1)} \tag{2.23}$$

Under the normal distribution these are minimum variance-unbiased estimators of variance components and then achieving OLS on the transformed data.

2.9.3.2. The Variance Component Modules Estimators

Looking back at the ancient history of estimating variance components, in (1956) Schaffer refers to a work as being “very explicit use of a variance-components model” for the one-way layout by Airy (1861) and each year after the other the opinions become more clearly and in a wider range applicable.

In a situation in which the noise or in another word the true disturbance cannot be identified so it can be unnecessary to use unbiased estimator of the variance components.so on dealing with this implementation gives a more efficient results as we are interested in using a large data set from population, so far many previously studied suggestions are available in order to deal with such impossible estimates as shown in equations (2.22) and (2.23).

One of the suggestions is by Marc Nerlove (1971) for estimators of the variance modules that is by use of ($\hat{\mu}_i$) that are the estimates of dummy coefficients from least square dummy variable (LSDV) of ($\sigma^2{}_v$)is estimated from dividing within sums of squares by (NT) in lack of any change in the degree of freedom:

$$\sigma_\mu{}^2=\left(\frac{(\sum_{i=1}^{N}(\hat{\mu}_i-\bar{\hat{\mu}})^2)}{(N-1)}\right)$$

A simple suggestion came out even with less effects by T D Wallace and Ashiq Hussain that refers to replacing the true disturbance by ordinary least squares residuals that is for random effects model the variance component estimates are consistent and unbiased, in specific problem, these estimates are easily computable and likely to be advisable.

Another suggestion is by following regressions by consistency of mean square errors of the regressions to obtain the estimates of the variance components suggested by Swamy-Arora (1972) that is uses residuals from the within regression (fixed effect) which results in obtaining the following;

$$\hat{\bar{\sigma}}_v{}^2=\frac{(yQy-yQX\acute{(}X\acute{Q}X\acute{)}^{-1}X\acute{Q}y)}{(N(T-1)-K)} \qquad (2.24)$$

Following by Between (Means) regressions across time:

$$\bar{y}_{i.}=\alpha+\bar{X}'_{i.}\beta+\bar{u}_{i.} \qquad (2.25)$$

$$i=1,2,\ldots.,N$$

normally the averages are repeated (T) times for each individual that results in (NT) observations but from the above equation (N) observations are exist, so we must treat this caution by running a cross section, again we come across verifying

($\sigma^2{}_1$.) this yields as (s^2), as a result of two step regression transformation on the error we get zero mean and the variance-covariance matrix

$$\begin{pmatrix} \sigma^2{}_v Q & 0 \\ 0 & \sigma^2{}_1 P \end{pmatrix}.$$

So on, any transformation results in general least square formulation on this system affecting equally on the basic analysis of panel data and estimators by GLS is an average weighted matrix of Within and Between estimates that are weighted by the consistent variance.

2.9.3.3. Advantages and Disadvantages of Random Effects Model

- **Advantages**
  - The multi dimensionality are moderately allowed even for very large models may be almost estimated
  - Permits inclusion of variables (like GDP) that vary in the same dimension as the random effects.
  - It is possible to estimate shrunken residuals.
  - -In the Between-studies variance is substantial (and statistically significant) instantly random effects models is the best one that deals with it.

- **Disadvantages**
  - Random effects models may estimate the effects of time-invariant variables, but the estimates may be biased because we are not controlling for omitted variables.
  - Random effects obligate that unobserved heterogeneity yields some probability constraints.
  - In random effects model there may be no specific determination of the distribution or briefing any of characteristics.

- In practical studies as one does a comparison between fixed and random effects models, in most of the times it can be clear that the efficiency of fixed effects models is more powerful.

2.9.3.4. Hausman's Test

Hausman's test is measured as a strong specification procedure to distinguish between fixed effects model and random effects model in the panel data analysis.

J. A. Hausman (1978) discovered significant different results in fixed effects specification as compared to random effects model but it is not so large. If the model is specified correctly and the assumption of random effects fails, then the fixed effects estimator stays consistent, but the random effects estimator is inconsistent, in this situation the difference may be large. Hausman's formal results can apply not only to linear models, but to maximum likelihood methods generally.

The assumption of the Hausman's test in the error component models is $E(u_{it}/X_{it} = 0)$ that is the sufficient factor to tell us the disturbances include individuals invariant effects (μ_i) which can be correlated with the independent variables, here $(E(u_{it}/X_{it}) \neq 0)$ and the general least square estimator $(\widehat{\beta}_{GLS})$ will be biased and inconsistent of (β), even the within transformation removes individuals effects and the within estimator $(\widetilde{\beta}_{within})$ becomes unbiased and consistent for (β).

Suggestion of Hausman's test is differentiating $(\widehat{\beta}_{GLS})$ and $(\widetilde{\beta}_{within})$ both of which are consistent under the null hypothesis H_0: $E(u_{it}/X_{it}) = 0$ but with different probability restricts if (H_0) is rejected.

Actually $(\widetilde{\beta}_{within})$ is consistent either (H_0) is true or not while $(\widehat{\beta}_{GLS})$ is a best linear unbiased estimator consistent under (H_0) but inconsistent if (H_0) is false.

A natural statistic test will be applied that says $(\hat{q}_1 = \widehat{\beta}_{GLS} - \widetilde{\beta}_{within})$ under (H_0) and $cov\left(\hat{q}_1, \widehat{\beta}_{GLS}\right) = 0$.

Appling $\widehat{\beta}_{GLS} - \beta = (\acute{X}\ \Omega^{-1}X)^{-1}\acute{X}\ \Omega^{-1}u$ and $\widetilde{\beta}_{within} - \beta = (\acute{X}QX)^{-1}\acute{X}Qu$

We get $E(\hat{q}_1) = 0$

So with $cov\left(\hat{q}_1, \widehat{\beta}_{GLS}\right) = 0$:

$$var(\hat{q}_1) = \sigma^2{}_v(XQX\acute{)}^{-1} - (\acute{X}\ \Omega^{-1}X)^{-1} \quad (2.26)$$

Therefore the Hausman's test is:

$$m_1 = \hat{q}_1\acute{\hat{q}}_1(var(\hat{q}_1))^{-1} \quad (2.27)$$

Simply Hausman's test is generalized for testing that each (X_{it}) uncorrelated with (μ_i), here we determine (NT) by the coefficient of the regression of the general least square residuals on $(\tilde{X}_{it})$ and $(\acute{X}_{i1}, \ldots\ldots, \acute{X}_{iT}\acute{)}$ that is recommended for testing joint hypothesis of exogeniety of the regressors and steady of the parameters over time otherwise all of the parameters even thorough exogeniety, they can be inconsistent.

2.9.4 Mixed Effects Models

In another name (Mixed Error Component Model), such models are presented as a special case of panel data modeling which is a mixture of fixed and random coefficients model. It is preferably in use because of its dialed missing values. To facilitate our discuses about such modeling, we must assume that only time invariant cross sectional heterogeneity appears.

Let's suppose that every cross sectional unit assumed to be disparate, thus;

$$y_{it} = \sum_{k=1}^{K}\beta_{ki}x_{kit} + \sum_{\ell=1}^{m}\gamma_{\ell it}w_{\ell it} + u_{it} \quad (2.28)$$

i=1,2,….,N

t=1,2,….,T

Where:

(x_{it}): (K ×1) vector

(w_{it}): (m×1) vector

Those for each of (xit) and (wit) are from explanatory variables that are independent of the error of the equation (u_{it}). By collecting the (NT) observations, we get:

$$y = X\beta + W\gamma + u \tag{2.29}$$

The equations (2.28) and (2.29) presume a relation of different behavioral equation for every cross sectional unit, here the advantage of pooling comes out which is only for putting the above equation (2.29) unrelated regression to maintain the efficiency of the estimates of the individual behavioral equation.

Stimulating for a mixed model, in a condition that there exist fundamental difference among cross sectional units we may be able to draw inferences in certain population descriptions through the application of inferential constraints on the coefficients of (x_{it}) and (w_{it}).

Here two kinds of restrictions exist;

1st: The coefficients of (x_{it}) are supposed to be an issue of stochastic restrictions of the form:

$$\beta = A_1\bar{\beta} + \alpha \tag{2.30}$$

Where:

(A_1): is an (NK×L) matrix with known elements

($\acute{\beta}$): is an (L ×1) vector of constants

(α) is assumed to be normally distributed, random variables with (0) mean and non-singular constant covariance matrix (C) and independent of (x_{it}).

2nd: The coefficients of (w_{it}) are assumed to be subject to:

$$\gamma = A_2\acute{\gamma} \tag{2.31}$$

Where:

(A_2): (Nm×n) matrix with known elements

$(\acute{\gamma})$: (n×1) vector of constants

As (A_2) is known we substitute (2.31) into (2.29) model as:

$$y = X\beta + \widetilde{W}\acute{\gamma} + u \tag{2.32}$$

Subject to (2.30) where ($\widetilde{W} = WA_2$).

The form (2.29) can be agreed for many possible fixed parameter configurations. Various linear panel data models with unobserved individual specific but time-invariant heterogeneity can be assumed as special cases of the models (2.29) and (2.31), these involve:

1- For all cross sectional units there is the same model
2- Different cross sectional units with different models.
3- (Variable intercept model) Interindividual differences remain constant over time if conditional on the observed exogenous variables.
4- Random coefficient model.
5- (Error Component Models) when the individual's specific effects, time invariant omitted variables are entertained as random variable.

2.10. Auto-Correlation

In a statistical view this issue also known as serial correlation, mentions about the degree of correlation of the same variables between two succeeding time intervals and it is a common problem in regression analysis that involves time series data. So on one of the assumptions of the classical linear regression analysis is non-correlating the error terms (u_i) but if the error terms correlated then result in inefficiency that means they are not best linear unbiased estimator any more but they are still consistent and distributed normally, but in heterogeneity cases the procedure of hypothesis testing may not be trusted because the standard error estimates are not well defined, so one must find out if there exist auto correlation and try to perceive another estimating procedure .

For detecting auto correlation the most effective and widely used test is Durbin-Watson d test, which is the sum of squared difference of successive residuals over the outstanding sum of squares:

$$\mathrm{DW}=\frac{\sum_{i=2}^{n}(e_t-e_{t-1})^2}{\sum_{i=1}^{n}e_t^2}$$

For panel data this test is generalized as when $e_{i,t}$ is the residual from an ordinary least square regression each cross sectional unit *i*, associated with the observation in panel *i* at time *t*, then the test is:

$$\mathrm{D}_{pd}=\frac{\sum_{i=1}^{N}\sum_{t=2}^{T}(e_{i,t}-e_{i,t-1})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}^2}$$

We must mention that by losing one observation in accounting for successive difference of residuals so the degree of freedom become (n-1) and the D value always lies between (0 and 4). Follows by removing any auto-correlation problem in the model, if exists, that there are many ways to use which depends on

the value of autocorrelation even by ‘generalized least squares estimation, iterative method and tow stage procedure ordinary least square estimations’.

2.11. Heteroscedasticity

Heteroscedasticity means differing variances that can be mentioned as regression disturbances whose variances are not constant and differs across observations which are usually unbiased and dependent.

There are major reasons for occurrence of Heteroscedasticity:

- Due to the presence of an outlier.
- In cross sectional data when small, medium and large organizations sampled together.
- If major variables eliminated from the model.
- Due to the presence of symmetrical and asymmetrical curves of the regressors included in the model.
- Due to false data transformation and incorrect functional form.

For detection of these cases, there are many test and ways that have been proved. first of all one must denote that the OLS estimates are constructed such that there will be no correlation between errors and the explanatory variables in order to notice if larger values of the errors are associated with some function of the explanatory variable then the test will consider if any of the coefficients in the model are significant, alternatively we conduct a similar analysis using the absolute value of error rather than the square since the logic of them is same. Finally one can use one of these corrections either the usage of a different specification for the model, applying a weighted least squares estimation method or Heteroscedasticity-consistent standard error.

3. APPLICATIONS AND ANALYSIS

3.1. Introduction

We are about to study practically on our data but before preceding analysis, should notice that the panel data in this study is a balanced panel data because the number of time-observations is the same for each individual. The data here also called long panel data, since in long panel the number of time period is greater than the number of cross section. Supposed one wants to estimate models and make prediction of wheat production by giving the area of agriculture and rainfall amount as independent variables and testing models for choosing the best model.

3.2. Data Description

The data of this study includes annual wheat production in tons and areas in decare that used with year on year and rain fall by milliliter in the countries of Turkey, Iraq and Iran, during the years (1991 to 2019). This contains (87) observations, in which these data are time series data and cross section data which also means the combined time series and cross section. Panel data was taken for Iraq country from “The Ministry of Agriculture” (Agricultural Statistics Office) and “Meteorological Office of Weather”, for Turkey country from “The Ministry of Agriculture and Forestry” and “Turkish State Meteorological Service” and for Iran country from “Deputy of Planning and Economic Affairs” (Information and Communication Technology Center) and “Water Resources Management Company” (Statistic and Reporting System). The dependent variable is wheat production and the independent variables are areas that used for producing wheat and the rain fall amount.

3.3. Variables of the Study

3.3.1. Dependent Variable

The dependent variable is the wheat production for each year and measured in tons.

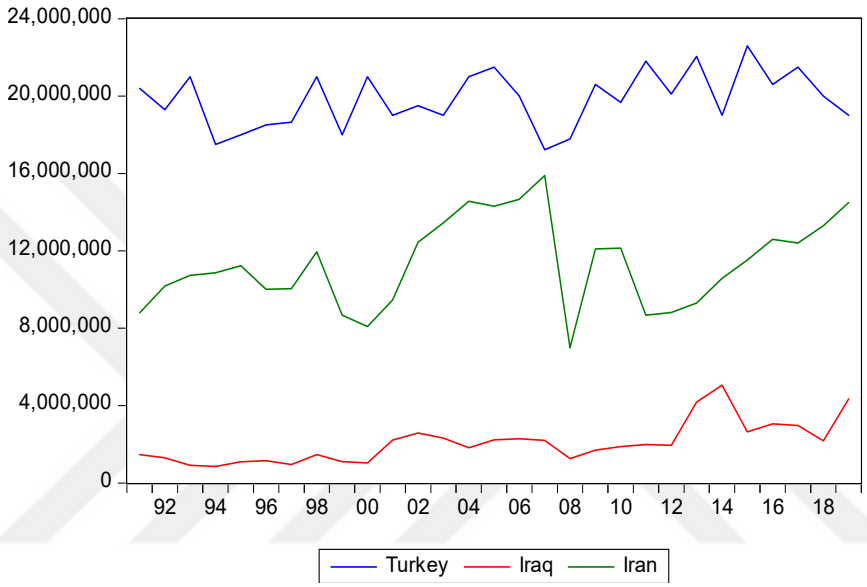

Figure 3.1. The graph of wheat production in the countries Turkey, Iraq and Iran.

Figure 3.1., Explains wheat production in the countries Turkey, Iraq and Iran from 1991-2019, it is clear that the maximum wheat production of Turkey, Iraq and Iran countries are (22,600,000; 5,055,000 and 15,886,090) tons respectively, and it is clear that the wheat production of Turkey country is greater than the wheat production of other countries for the year (2015), in which the rainfall for the same year was (557) milliliters and the area was (78,668,874) decares.

3.3.2. Independent Variables

We have two independent variables; these are the areas that were used for producing wheat, measured by decare and the rainfall amount measured by milliliter.

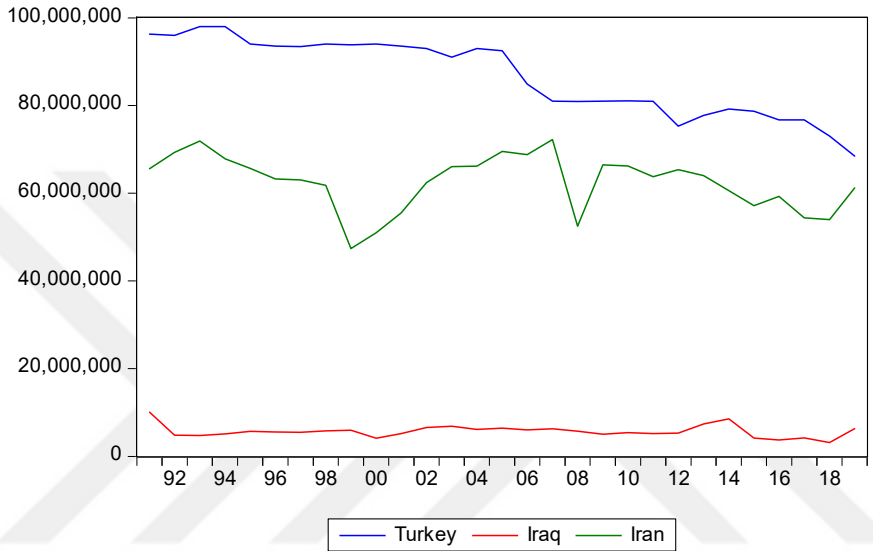

Figure 3.2. Graph of the area for producing wheat for each of the countries Turkey, Iraq and Iran.

Figure 3.2., explains the area of wheat production in Turkey, Iraq and Iran from 1991-2019. The maximum areas of them are 98,000,000; 10,069,000, and 72,223,110 decares respectively, also it is clear that Turkey country has the maximum average of area for wheat production that the product for the same year is equal to 21,000,000 tons.

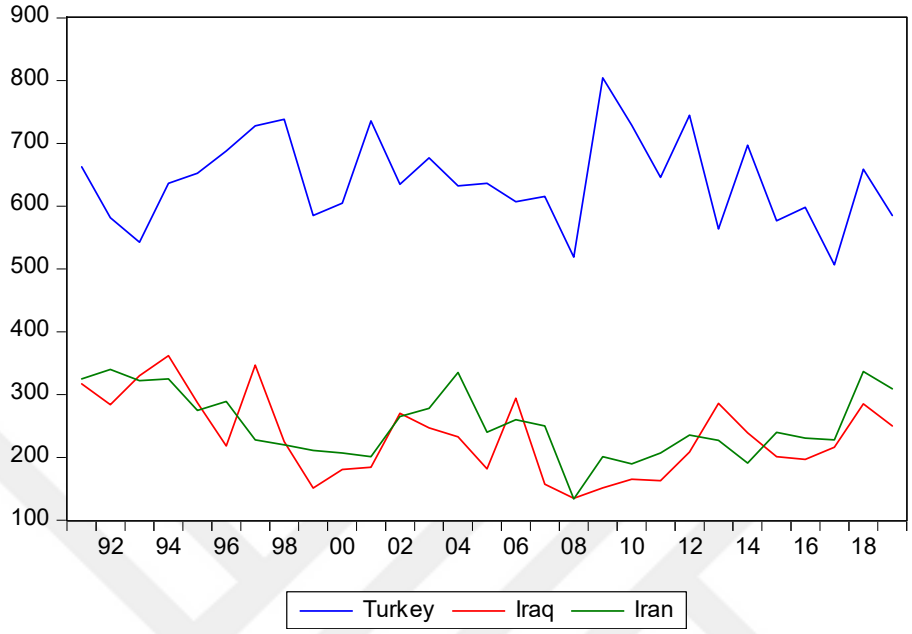

Figure 3.3. Graphs of rainfall for the countries Turkey, Iraq and Iran.

Figure 3.3., explains the rainfall in Turkey, Iraq and Iran from 1991-2019, the rainfall of each of the countries are represented as the maximum rainfall amount of them, that are (804.2, 361.9, 340) milliliters, respectively. So on, Turkey country has the maximum average of rainfall.

3.4. Application

We are about to follow steps in order to achieve systematic analysis of panel data from our collected data that we've mentioned before, for panel data analysis we used SPSS and Eviews9 software's. Of course after standardizing all of the variables under the study we attained the following points:

3.4.1. Heteroscedasticity Test

In this study we have two types of data cross-section and time series data, for this reason we must test data for heteroscedasticity and autocorrelation

problem. Using (Goldfield- Quandt Test) to test heteroscedasticity problem where the data divided into two groups, for each group the results of (Mse_1) and (Mse_2) were equal to (0.053, 0.034) in turn, and then applying F-test statistic was equal to (0.64). This can further be seen when the F-test statistic was less than (F_{table}) which is equal to (1.64) that means we can accept the null hypothesis and not having the heteroscedasticity problem.

3.4.2. Pooled Model

Estimating the pooled regression as initial model estimation and testing for autocorrelation problem using Durbin Watson test, the results are shown in table 3.1.

$$\hat{y}_{it} = 0.746682X_1 + 0.290881X_2 \quad (3.1)$$

Table 3.1. Shows the summary of the result for the pooled model

Model	Coefficient		P-value	SSE	df	MSE	Akaike	Adj R^2	D.W
Pooled OLS	b_1	0.746682	0.0000	3.74	84	0.045	-0.24	0.95	0.72
	b_2	0.290881	0.0000						

From the Table 3.1., the column of the coefficients represents the estimated parameters of the area per decares and the rainfall per milliliters which are equal to (0.746682 and 0.290881), respectively. This means, the harvests of wheat production are increased by (0.746682) per decare of area and it is increased by (0.290881) per mm of rainfall as it is obvious from the P-value of the estimators, which is less than (0.05), and it implies that both estimators are statistically significant.

In another words, the null hypothesis that says ($\beta_1 = \beta_2 = 0$) is rejected. The coefficient of determination is (%95) that means both independent variables are capable of explaining %95 of the dependent variable, but the other %5 is related to unknown information that this study could not obtain them. The appraisal measurements (Mse and AIC) are equal to (0.045 and -0.24) incidentally, as they should be used for comparison measures among the candidate models to get the best postulated model that fits the data under study. The predicting values after running pooled model are shown in the Figure 4., Below:

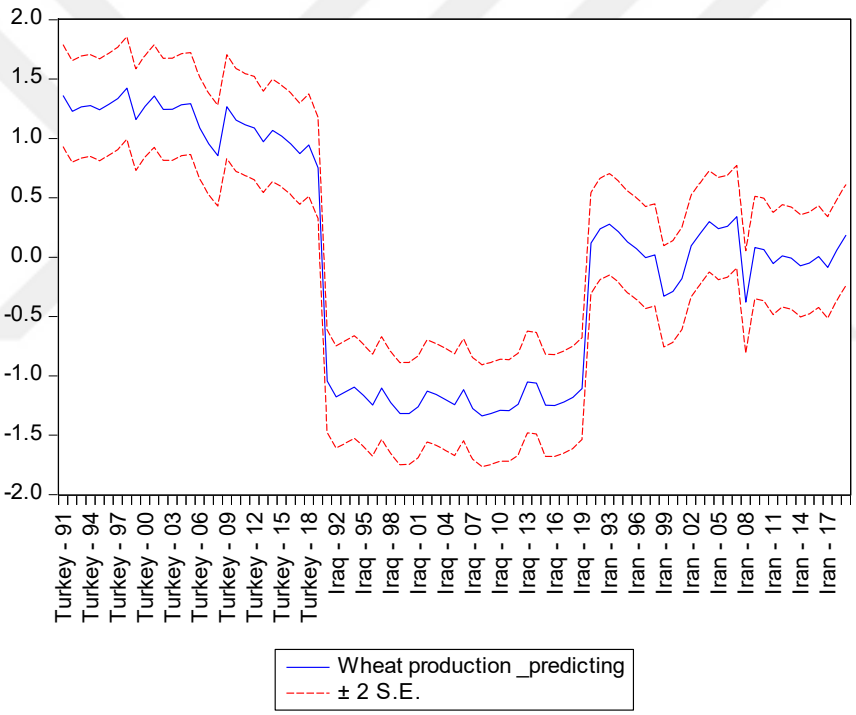

Figure 3.4. Graph of predicting wheat production for the pooled OLS model.

Figure 3.4., explains predicting wheat production using pooled OLS Model, by Review on the figure, it is clear that predicted values fell between the upper bound and lower bound.

The hypothesis of autocorrelation problem has been tested depending on the Durbin-Watson test which is clear from the table 3.1., in which the calculated D.W test value is (0.72) and it is less than the ($d_u = 1.553$) from D.W table as it implies that the model suffers from A.C problem. Otherwise, one can say that, there exist a systematic pattern in the residuals of the model, therefore one should remove A.C problem and then rerun the pooled model again. After removing the A.C problem by using iterative method the results of the Pooled model is as shown in Table 3.2:

Table 3.2. Shows the summary of the pooled model after removing AC problem

Model	Coefficient		P-value	SSE	df	MSE	Akaike	Adj R^2	D.W
Pooled OLS	b_1 b_2	0.834694 0.43189	0.000 0.000	2.51	84	0.03	-0.63	0.89	2.1

As it is clear that the calculated D.W test value is (2.1) and it is fallen between d_u and 4-d_u that tells the model which does not suffer from AC problem anymore and noticed that the coefficient of determination (Adj R^2) is decreased by (%6) that tells the statisticians which prefers a model with lower (Adj R^2) but without suffering from any econometrics problem.

The predicting values of pooled model after removing AC problem are shown in the graph in Figure 3.5.

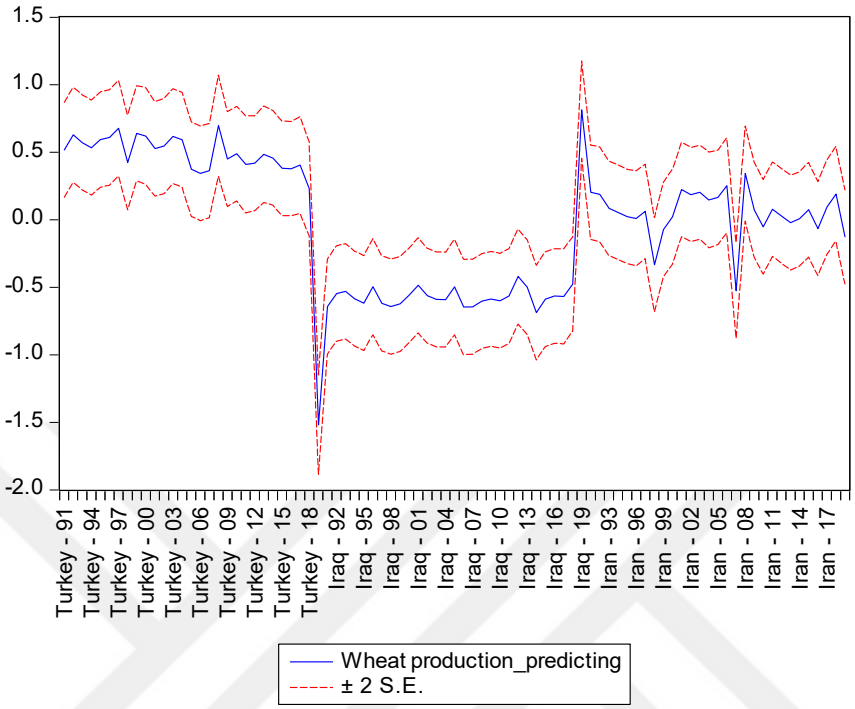

Figure 3.5. Graph of predicting wheat production for the pooled model after removing AC.

Figure 3.5. explains predicting wheat production, using pooled model after removing AC problem, from the review on the figure, it is clear that predicted values fell between the upper bound and lower bound.

3.4.3. Fixed Effects Model Estimation (fixed cross section)

Estimating the fixed effect model fixed for cross section as initial model and testing for autocorrelation problems by using Durbin Watson test, its results are shown in the table 3.3. :

$$\hat{y}_{it} = 0.508139X_1 + 0.145357X_2 \quad (3.2)$$

Table 3.3. Shows the summary result for fixed effect model fix for cross section

Model	coefficient		P-value	SSE	df	MSE	Akaike	Adj R^2	DW
Fix (cross fix)	b_1	0.508139	0.0004	3.41	82	0.04	-0.28	0.95	0.89
	b_2	0.145357	0.0515						

From Table 3.3., the estimated parameters of area and rainfall that are shown in the column of coefficients are equals to (0.508139 and 0.145357) respectively, so the yields of wheat are increased by (0.508139) per decare of area and it is increased by (0.145357) per mm of rainfall .The adjusted coefficient of determination ($AdjR^2$) is (%95) that means both explanatory variables can explain %95 of the response variable. The comparison measurements (Mse and AIC) are equal to (0.04 and -0.28) incidentally, in which they should be used as a comparison measures amongst the candidate models to get the best postulated model which fit the data under study.

The predicting values after running fixed effect model fixed for cross section are shown in the graph, in Figure 6.:

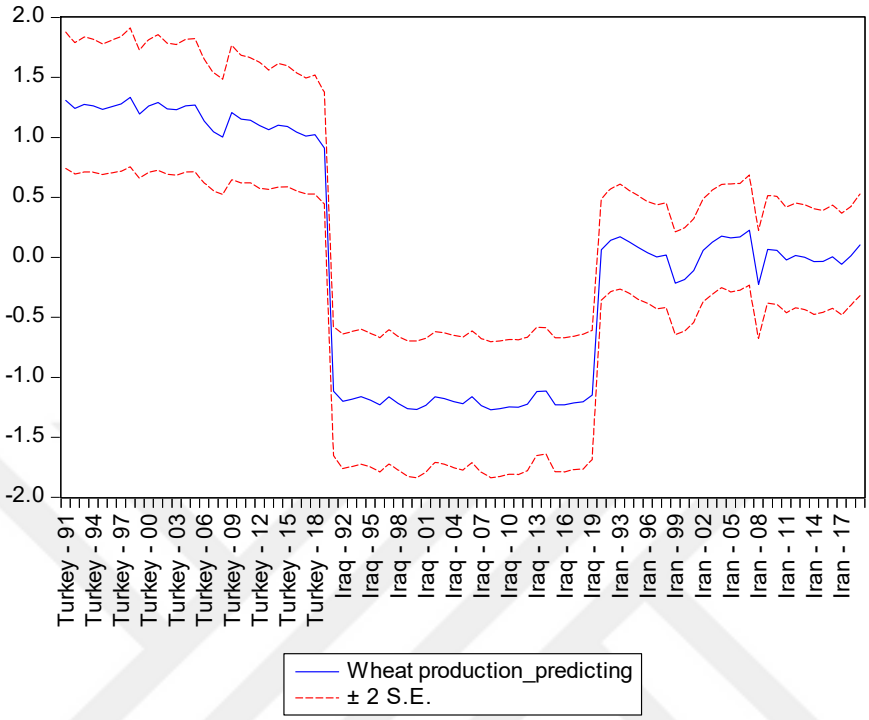

Figure 3.6. Graphs of predicting wheat production for fixed effect model fixed for cross section.

Figure 3.6. explains predicting of wheat production using fixed effects model, fix for cross section, in the figure 6. the predict values fell between the upper bound and lower bound.

The hypothesis of autocorrelation problem has been tested depending on Durbin-Watson test value is (0.89) and it is less than the ($d_u = 1.553$) from D.W table it implies that the model suffers from AC problem or there exists a systematic pattern in the residuals of the model, therefore we should remove AC problem and then rerun the fixed model again.

After removing the AC problem by using iterative method the results of the fixed effect model fixed for cross was as shown in the Table 3.4.

Table 3.4. The summary result for fixed effect model fix for cross section after removing AC.

Model	Coefficient		P-value	SSE	df	MSE	Akaike	Adj R^2	D.W
Fix(cros fix)	b_1	0.855714	0.0000	2.51	82	0.03	-0.59	0.90	1.89
	b_2	0.178642	0.0043						

Table 3.4. Shows the summary of fixed effects model fixed for cross section after removing AC problem, now the calculated D.W test value is (1.89) and it is fallen between d_u and 4-d_u, tells us the model no more including autocorrelation problem and noticed that the adjusted determination coefficient is decreased by %5, likely we prefer a model with low R^2 but without suffering from any econometrics problem.

The predicting values for fixed effect model, fixed for cross section, after removing AC problem are shown in the graph in Figure 7.

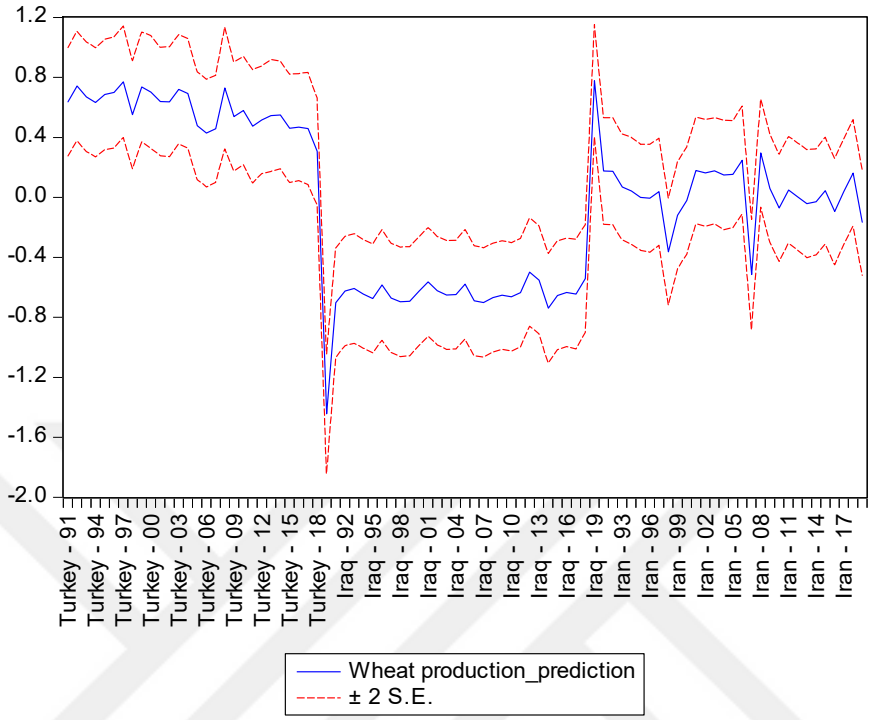

Figure 3.7. Graph of predicting wheat production for fixed effect model fixed for cross section after removing AC problem.

Figure 3.7. explains predicting of wheat production using fixed effects model fixed for cross section after removing AC problem, shows that the predicted values are fallen between the upper bound and lower bound.

3.4.4. Fixed Effects Model Estimation (fixed cross section and period)

Estimating the fixed effect model fixed for cross section and period as initial model and testing for autocorrelation problem using Durbin Watson test, its results are shown in the Table 3.5.

$$\hat{y}_{it} = 0.747737X_1 + 0.219803X_2 \quad (3.3)$$

Table 3.5. The summary result for fixed effect model, fixed for cross section and period.

Model	Coefficient		P-value	SSE	df	MSE	Akaike	Adj R^2	D.W
Fix (both fix)	b_1	0.747737	0.0000	1.43	82	0.02	-0.51	0.97	1.23
	b_2	0.219803	0.0052						

In this table the estimated parameters of area and rainfall are equals to (0.747737and 0.219803) respectively, shows the yields of wheat are increased by (0.747737) per decares of area and it is increased by (0.219803) per mm of rainfall. The P-value of the estimators is less than (0.05) that implies the significance of estimators following by the hypothesis of ($\beta_1 = \beta_2 = 0$) rejection. The adjusted coefficient of determination ($AdjR^2$) is (%97) that means both explanatory variables are capable of explaining %97 of the response variable. The comparison measurements (Mse and AIC) are equal to (0.02 and -0.51) respectively.

The predicting values after running fixed effect model fixed for cross section and period are shown in the graph, in the Figure 8.

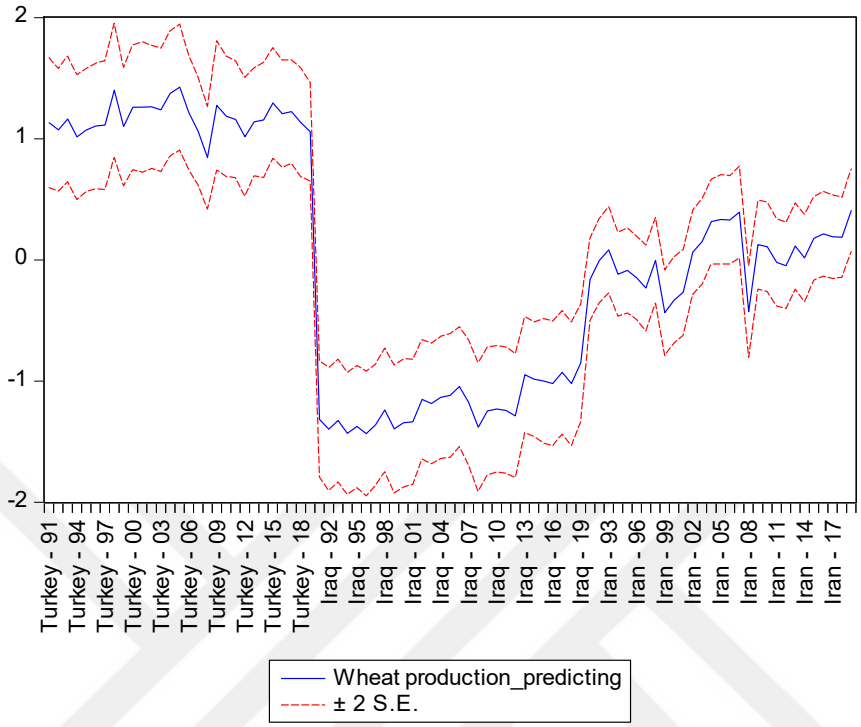

Figure 3.8. Graph of predicting wheat production for fixed effect model fixed for cross section and period.

Figure 3.8. explains predicting of wheat production using fixed model for cross section and period, from the figure 8., the predict values are fallen between the upper and lower bounds.

The calculated D.W test value is (1.23) which is less than the ($d_u = 1.553$) as stated in D.W table, accordingly the model suffers from AC problem.

After removing the AC problem by using iterative method, the results of the fixed effect model fixed for cross section and period obtained as shown in Table 3.6.

Table 3.6. The summary result for fixed effect model fixed for cross section and period after removing AC.

Model	Coefficient		P-value	SSE	df	MSE	Akaike	Adj R^2	D.W
Fix(both fix)	b_1 b_2	0.800735 0.182565	0.0000 0.0077	1.277	82	0.0155	-0.62	0.94	2.07

Table 3.6. Shows the summary of fixed effect model, fixed for cross section and period, after removing AC problem, from the table 3.6., the calculated D.W test value is (2.07) and it is fallen between d_u and 4-d_u that tells the model not including any AC problem anymore, and we noticed that the determination coefficient R^2 is decreased by %3, so on a small decreasing in R^2 but without existence of econometrics problem.

The predicting values for fixed effect model fix for cross section and period after removing AC problem are shown in the graph in Figure 9.

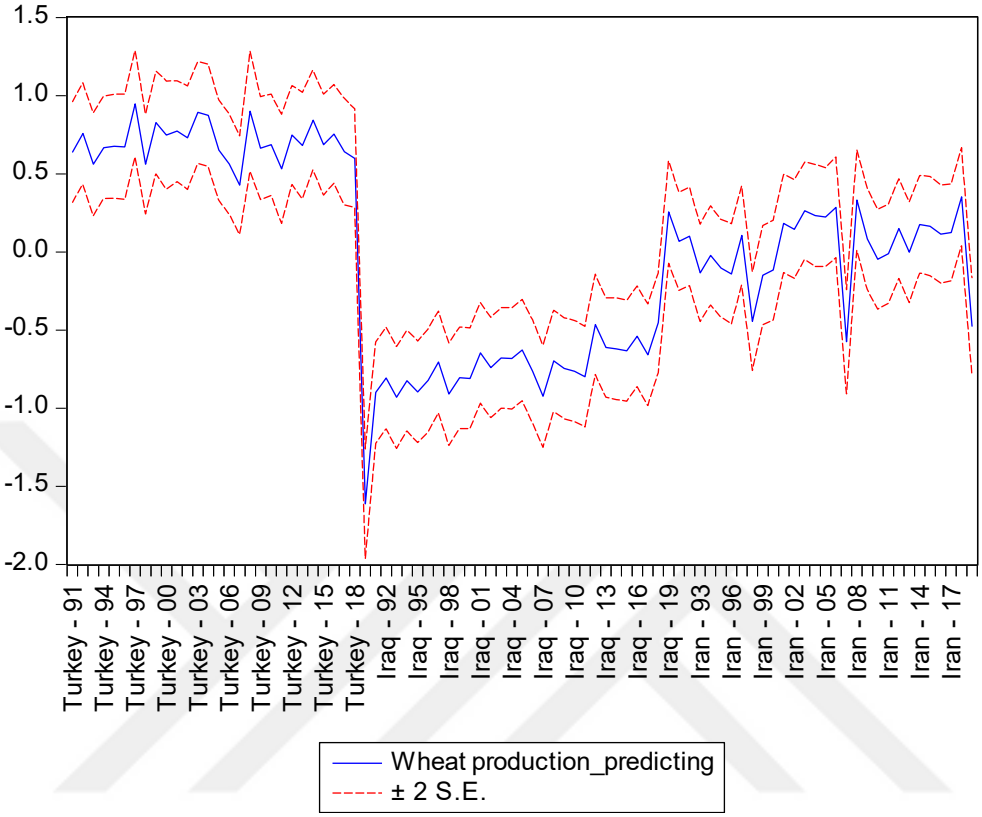

Figure 3.9. Graph of predicting wheat production for fixed effect model, fixed for cross section and period, after removing AC problem.

Figure 3.9., explains predicting of wheat production using fixed effect model fixed for cross section and period after removing AC problem. It shows that the predicted values are fallen between the upper bound and lower bound.

3.4.5. Random effects model estimation (random for cross section)

Estimating the random effect model random for cross section as initial model using Swamy and Arora estimators of variance components and testing for autocorrelation problem using Durbin Watson test, the results are shown in Table 3.7.

$$\hat{y}_{it} = 0.746018X_1 + 0.2371767X_2 \quad (3.4)$$

Table 3.7. The summary result for random effect model random for cross section.

Model	Coefficient		P-value	SSE	df	MSE	Adj R^2	D.W
Random(cross random)	b_1	0.746018	0.000	3.669	84	0.04	0.76	0.69
	b_2	0.237176	0.001					

Table 3.7. Shows the summary result for random effect model, random for cross section, the table 3.7 shows rejecting the null hypothesis that says ($\beta_1 = \beta_2 = 0$), since the p value of estimates is less than (0.05). Each of estimated parameters, the area equal to 0.746018 and rainfall equal to 0.237176, that shows increasing in the yields of wheat by (0.746018) per decare of area and (0.237176) per mm of rainfall. The adjusted coefficient of determination ($Adj\ R^2$) is (%76) that means both explanatory variables are capable of explaining %76 of the response variable but the other %24 is related to an unknown information that are not achieved, with Mse is equal to (0.04).

The predicting values after running random effect model, random for cross section, are shown in the graphs in the Figure 10.

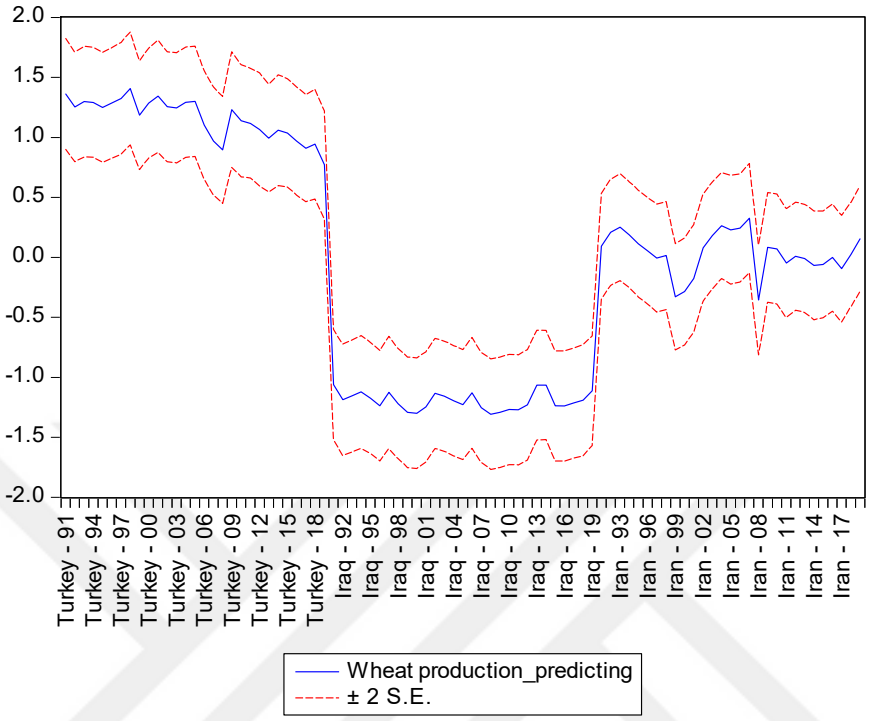

Figure 3.10. Graph of predicting wheat production for random effect model, random for cross section.

Figure 3.10., explains predicting of wheat production using RE model random for cross section. From the figure 10, the predict values lie between the upper bound and the lower bound.

We tested the hypothesis of autocorrelation problem depending on Durbin-Watson test and the calculated D.W test value is (0.69) and it is less than the ($d_u = 1.553$) from D.W table that determines existing of AC problem or there exists a systematic pattern in the residuals of the model, therefore we should remove AC problem and then rerun the random model again. After removing AC problem by using iterative method and re-running the random effect model using Wallace_Hussain estimator of variance components, the results of the random effect model are shown in Table 3.8.

Table 3.8. The summary of random effect model after removing AC problem

Model	coefficient		P-value	SSE	df	MSE	Adj R^2	D.W
Random(cross random)	b_1	0.852533	0.0000	2.48	84	0.03	0.86	2.15
	b_2	0.198543	0.0001					

The calculated D.W test value is (2.15) that mean no AC problem anymore with increasing in (R^2) by %10 that shows more efficiency tells us surviving from any econometric problem. The predicting values after running random effect model are shown in the graph in Figure 11.

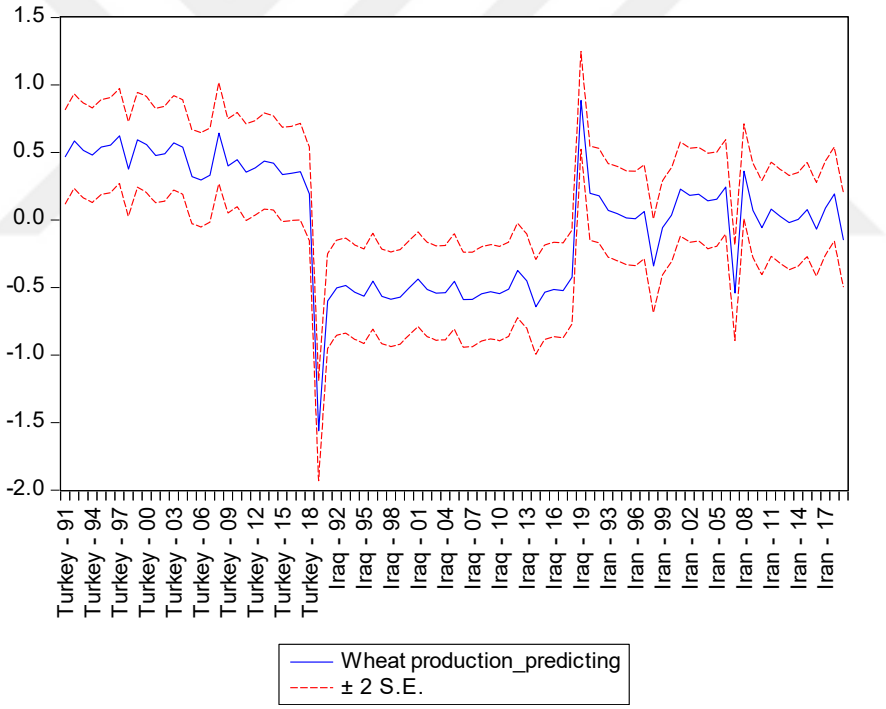

Figure 3.11. Graph of predicting wheat production for random effect model after removing AC problem

Figure 3.11. explains predicting of wheat production using RE model random for cross section after removing AC. From the figure 11., it is clear that the predicted values are lying between the upper and lower bounds.

3.4.6. Mixed Effects Model Estimation (random cross section and fix period)

Estimating the mixed effects model, random for cross section and fix for period, as initial model using Swamy and Arora estimator of variance components and testing for autocorrelation problem, using Durbin Watson test, the results are shown in Table 3.9. :

$$\hat{y}_{it} = 0.746036X_1 + 0.301547X_2 \quad (3.5)$$

Table 3.9. The summary result for mixed effects model, random for cross section and fix for period.

Model	Coefficient		P-value	SSE	df	MSE	Adj R^2	D.W
Mix (cross random period fix)	b_1 b_2	0.746036 0.301547	0.000 0.000	1.47	84	0.02	0.97	1.34

Table 3.9., clarifies rejecting the null hypothesis that says ($\beta_1 = \beta_2 = 0$) since the P-value of the estimators are less than (0.05), which implies that both estimators are statistically significant. The estimated parameters of area and rainfall are equal to (0.746036 and 0.301547) respectively, that means the yields of wheat are increased by (0.746036) per decare of the area and (0.301547) per mm of rainfall. The (AdjR^2) is (%97) that means both explanatory variables are capable of explaining %97 of the response variables, but the other %3 is remained unknown. The comparison measurement (Mse) is equal to (0.02) which should be used as comparison measures among the candidate models.

The predicting values after running mixed effects model, random for cross section and fix for period, are shown in the graph in Figure 12.



Figure 3.12. Graph of predicting wheat production for mixed effects model, random for cross section and fixed for period

Figure 3.12., explains predicting of wheat production using mixed effects model, random for cross section and fix for period. The calculated D.W test value is (1.34) which is less than the ($d_u = 1.553$). From D.W table it implies that the model suffers from AC problem, or one can say that there exists a systematic pattern in the residuals of the model, therefore one should remove AC problem and then rerun the mix model again. After removing the AC problem by using iterative method, and rerunning the mixed effects model using Wallace and Hussain estimator of component variances, the results of the mixed effects model are shown in Table 3.10.

Table 3.10. The summary of mixed effects model, random for cross section and fixed for period, after removing AC.

Model	Coefficient		P-value	SSE	df	MSE	Adj R^2	D.W
Mix model (cross random)	b_1 b_2	0.772636 0.267790	0.0000 0.0000	1.328	84	0.0158	0.94	2.135

From the Table 3.10, the AC problem no more exists and become (2.135) but with less ($AdjR^2$) that decreased by %3 and it is preferable as no more suffering from any econometrics problem. The predicting values after running mixed effects model are shown in the graph in Figure 13. :



Figure 3.13. Graph of predicting wheat production for mixed effects model, random for cross section and fixed for period, after removing AC.

Figure 3.13. explains predicting of wheat production using mixed effects model, random for cross section and fix for period, after removing AC problem, shows lying of the predicted value between upper and lower bound.

3.4.7. Mixed Effects Model Estimation (random period and fix cross section)

Estimating the mix model period random and cross section fix as initial model using Swamy and Arora estimators of variance components and testing for autocorrelation problem using Durbin Watson test, the results are shown in the Table 3.11.

$$\hat{y}_{it} = 0.621181X_1 + 0.174654X_2 \quad (3.6)$$

Table 3.11. The summary result for mixed effects model, period random and cross section fix.

Model	Coefficient		P-value	SSE	df	MSE	Adj R^2	D.W
Mix(cross fix period random)	b_1 b_2	0.621181 0.174654	0.000 0.0108	2.26	84	0.03	0.97	0.99

From table 3.11., the estimated parameters of area and rainfall which are equal to (0.621181 and 0.174654) respectively, means increasing by (0.621181) per decare of area and it is increased by (0.174654) per mm of rainfall in the wheat yields, so on both estimators are statistically significant and the null hypothesis that says ($\beta_1 = \beta_2 = 0$) is rejected. Actually both explanatory variables are capable of explaining %97 of the response variable. The mean square error equal to (0.03) which would be used as a comparison measures among the candidate models to get the best model that fit the data under study.

The predicting values after running mixed effects model, period random and cross section fix, are shown in the graphs in the Figure 14.



Figure 3.14. Graph of predicting wheat production for mixed effects model, period random and cross fix.

Figure 3.14., explains predicting of wheat production using mixed effects model, fix for cross section and random for period. It is clear that the predicted values fell between the upper bound and lower bound.

From the Table 3.11., we see that the calculated D.W test value is (0.99) which is less than the ($d_u = 1.553$). From D.W table, it implies that the model suffers from AC problem, or one can say that there exists a systematic pattern in the residuals of the model. Then the AC problem should be removed and then run the mixed effects model again.

After removing the AC problem by using iterative method and re-running the mixed effects model using Swamy and Arora estimators of component variances, the results of the mixed effects model, period random and cross section fix, are shown in the Table 3.12.

Table 3.12. Shows the summary of mixed effects model, period random and cross fix, after removing AC problem.

Model	Coefficient		P-value	SSE	df	MSE	Adj R^2	D.W
Mix (period random)	b_1	0.856186	0.0000	2.013	84	0.02	0.90	2.05
	b_2	0.165056	0.0047					

Likely the auto correlation problem has been removed, the value is (2.05) although the (Adj R^2) is decreased by %7 that fell from %97 to %90, but as econometric problem in not existing any more we may prefer the new obtained model. The predicting values after rerunning the mixed effects model, period random and cross section fix, are shown in the graph in Figure 15.



Figure 3.15. Graphs of predicting wheat production for mixed effects model, period random and cross fix, after removing AC.

Figure 3.15. explains predicting of wheat production using mixed effects model, fix for cross section and random for period, after removing AC problem.

Summing up the Figure 15., it is clear that the predicted values are fallen between the upper bound and lower bound.

3.4.8. Models Fitting Results for the Panel Data

Table 3.13. The models summary and comparisons

Model	Coefficient		P-value	SSE	MSE	R^2	AdjR^2	DW
Pooled OLS	b_1	0.834694	0.0000	2.51	0.03	0.89	0.89	2.1
	b_2	0.213189	0.0000					
Fix (cross fix)	b_1	0.855714	0.0000	2.51	0.03	0.90	0.90	1.89
	b_2	0.178642	0.0043					
Fix (both fix)	b_1	0.800735	0.0000	1.277	0.0155	0.96	0.94	2.07
	b_2	0.182565	0.0077					
Random(cross random)	b_1	0.852533	0.0000	2.48	0.03	0.86	0.86	2.15
	b_2	0.198543	0.0001					
Mix model (cross random)	b_1		0.0000	1.328	0.0158	0.96	0.94	2.135
	b_2	0.772636	0.0000					
		0.267790						
Mix (period random)	b_1	0.856186	0.0000	2.013	0.02	0.91	0.90	2.05
	b_2	0.165056	0.0047					

The Table 3.13. explains summary of the models and the comparison between models to choose which one is the best, we can see definitely that the best model is the fixed effect model (both fix) depending on value of (MSE and R^2) the fixed effect model has lowest MSE equal to (0.0155) and highest R^2 equal to (0.96).

3.4.9. ANOVA Table

After fitting the six models, it is needed to test each of them separately as comes from the estimation and testing of the panel data, for this purpose (ANOVA) table is calculated for each model.

In all models, testing the hypotheses:

H_0: The model is not significant.

H_1: The model is significant.

Table 3.14. The variance analysis of all the models

Models	Source	SS	d.f	MS	F-statistic test
Pooled model	SSR	20.56	2	10.28	342.6042
	SSE	2.514171	84	0.03	
	SST	23.074171	86		
Fixed(cross)	SSR	12.3	2	6.15	205
	SSE	2.513209	82	0.03	
	SST	14.813209	84		
Fix(both)	SSR	1.78	2	0.89	44.5
	SSE	1.277228	82	0.0155	
	SST	3.057228	84		
Random(cross)	SSR	16.88	2	8.44	281.33
	SSE	2.488670	84	0.03	
	SST	19.36867	86		
Mix(cross random)	SSR	2.22	2	1.11	55.5
	SSE	1.328746	84	0.0158	
	SST	3.548746	86		
Mix(period random)	SSR	8.68	2	4.34	217
	SSE	2.013365	84	0.02	
	SST	10.693365	86		

The value of F- calculated, which comes out from the ratio of mean square regression by mean square error is equal to (342.6042, 205, 44.5, 281.33, 55.5, 217) separately, the F-table value with (2, 84, 0.95) is equal to 3.11, which is far smaller than F-calculated of each model. This implies that the models are fit, so we reject the null hypothesis which says the model is not significant.

3.4.10. Comparison of the Models

After running all the models of panel data, we can make the comparison between models using two specification tests:

3.4.10.1. Restricted F Test

Using F-restricted test is to choose appropriate model between pooled model and fixed model.

H_0: pooled model is appropriate.
H_1: fixed effect model is appropriate.

The results of the test show that fixed effect model is an appropriate model where the ($F_{calculated}$) for fixed effect models is equal to (44.5), this is comparing to (F_{table}) ($F_{(N-1),N(T-1)-K}$) which is equal to ($F_{(2,82)}$=3.1) as it is clear one can see that ($F_{calculated}$) is greater than (F_{table}). This also means one can reject the null hypothesis that says the pooled model is an appropriate model.

3.4.10.2. Hausman's Test

Hausman test is the test, which decides the model, fixed effects model or random effects model as an appropriate model.

H_0: random effect model is appropriate.
H_1: fixed effect model is appropriate.

The result of Hausman test is shown in the Table 3.15.

Table 3.15. The result of the Hausman's test.

.

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	6.318842	2	0.0425

Cross-section random effects test comparisons:

Variable	Fixed	Random	Var(Diff.)	Prob.
X1_	0.508139	0.746018	0.014165	0.0456
X2_	0.145357	0.237176	0.002070	0.0436

As shown in the Table 3.15., we can discover a statistically significant p-value equals 0.04 which is less than 0.05 for that one may reject the null hypothesis which also says random effect model is an appropriate model. So the appropriate model is the fixed effects model.

4. CONCLUSION AND POSSIBLE RECOMMENDATIONS

Through the study of the topic under the thesis "**Panel Data Analysis and Modelling in the Field of Agriculture",** we found a number of conclusions and recommendations around it, mainly includes;

4.1. Conclusions

As a consequence of the results, we obtained from our practical work by applying regression analysis on our panel data, we maintain the following conclusions:

- Foremost pooled model for panel data cannot give an adequate performance although it does not include any significant econometrics problem, but the main behavior of panel data originally contains correlated error specially that is concerning with time domain behavior.
- Pooled regression model for panel data analysis seems to be ineffectual, since it cannot reveal every possible influence of the data.
- From the F – restricted test, one determines that the fixed effects model is a more appropriate model than the pooled model.
- From the Hausman's test that is to compare which model (fixed or random) is appropriate, then from the result of this test, it is concluded that the data is homogenous, so the fixed effect model is more appropriate because the null hypothesis that says random effect model is appropriate has been rejected.
- Among the six fitted models, the best model was fix effects model (both fix), that has been chosen depending on the mean square error which is equal to 0.0155, that is the model which will be used to predict wheat production.

- The fixed effects model mean square error result is close to that of the mixed effects model but here we can determine that the performance of the fix effects model is higher with $R^2 = 0.96$.
- The estimators of parameters for both explanatory variables are statistically significant in all six fitted models that imply the information in the data set under consideration is the sufficient information.
- The mean square error of fixed effects model equal to 0.0155 but the mean square error of mixed effects model (Random cross section) is equal to 0.0158 so even this small difference in a reliable reason for choosing fixed effects model.
- Since we have homogeneity of the panel data under consideration and the best performance of fixed effects model, then it is concluded that the three locations of countries, (Turkey, Iraq and Iran) generally have the same chance of wheat production as quantities.

4.2. Recommendations

This study presents some important point to recommend the stakeholders and ministry of agriculture of each of the countries of (Turkey, Iraq and Iran) to focus on:

- The places where the rainfall is low but with much agricultural land, the governments must make facilities for farmers and help them by organizing all over needs for performing the products all over the country.
- More studies are required for determination of the proportion of rainfall for necessity of wheat production.
- The stakeholders should motivate farmers to improve agriculture by reducing imported-production goods and staying a good marketing policy for internal production.

- The researchers may be recommended to try panel data analysis in order to get adequate results are more powerfully when the data is multidimensional by increasing in explanatory variables agricultural tools or the seeds quality also required for collecting data about these explanatory by the responsible stakeholders to make these kinds of research more precise and flexible.
- For future studies one can benefit from deeper model analysis including dynamic regression models which may be work more efficiently and precisely on the panel data.

REFERENCES

Ahmad, G., 2017. Empirical assessment of Climate Change on Agricultural Crops: Panel Data Analysis from Pakistan. Researchgate, 5(2), 59-68. Ministry of Planning and Reform, Pakistan.

Arellano, M., 2004. Panel Data Econometrics. Oxford University Press, New York, 244p.

Amemiya, T., 1971. The estimation of the variances in variance component model. International economic review, 12(1), 1-13.

Arellano, M. and Bonhomme, S., 2012. Identifying Distributional Characteristics in Random Coefficients Panel Data Models. The Review of Economic Studies. 79(3), 987-1020. Oxford University Press.

Ahmed, M. M., 2012. Market Structure and Performance of Bangladesh Banking Industry: A Panel Data Analysis. The Bangladesh Development Studies, 35(3), 1-18. Bangladesh Institute of Development Studies.

Arellano, A., 1987. Computing Robust Standard errors for within – group's estimators. Oxford, Bulletin of economics and statistics. 49(4), 431-434.

AL Ansari, N., 2014. Expected Future precipitation in general Iraq using LARS-WG Stochastic Weather Generator. In Engineering, 6: 948-959, Lulea university of Technology.

Biorn, E., 2017. Econometrics of Panel Data Methods and Applications. Oxford University Press, United Kingdom, 417p.

Baltagi, B. H., 2006. Panel Data Econometrics. The Theoretical Contributions and Empirical Applications. Elsevier B. V., Syracuse University, New York,380p.

Baraka, M. and Jabury, M., 2014. The Effect of Exchange Rate Regimes Nature on Inflation Analyze the Relationship by Using Panel Data. Journal of the performance of Algerian institutions. No:6.

Baez, A., 2014. A Panel Data analysis of FDI and Informal Labor Markets. In Research Institute of Applied Economics and Regional Quantitative Analysis Research Group.

Babel, B., Bomsdorf, E. and Schmidt, R., 2008. Forecasting German Mortality Using Panel Data Procedures. Journal of Population Economics, 21(3), 541- 555. Springer.

Bhole, L. M. and Mahakud, J., 2004. Behavior of Trade Credit: Time Series and Panel Data Analysis. Economic and Political Weekly, 39(1).

Canay, I. A., 2011. A Simple approach to Quantile regression for panel Data. Econometric Journal,14, 368-386.

Chen, X. and Guangqing Chi, G. S., 2012. Natural Beauty, Money and the Distribution of Talent: A Local-Level Panel Data Analysis. Population Research and Policy Review. 31(5), 665-681. Springer in cooperation with the Southern Demographic Association.

Davidson, R. and Mackinnon, J. G., 1993. Estimation and inference in econometrics. Oxford University Press, New York, 871p.

El-Hashash, A.,2017. Comparison of Variance components method for One Wa Random Effects Model in Cotton. In Asian Journal of Advances i Agricultural Research, 3(1), 1-9.

El Kharti, L., 2014. The Determinants of Financial Performance of Microfin Institutions in Morocco: A Panel Data Analysis. Savings and Development, 38(1). 27-44. University of Bergamo.

Frees, E. W., 2004. Longitudinal and Panel Data: Analysis and Applications in the Social Sciences. Cambridge University Press, New York, 485p.

Gurka, M., Kelly G., Edward L., 2012. Fixed and Random Effects Models. In Wiley Interdisciplinary Reviews Computational Statistics.

Gujarati, D., 2011. Econometrics by example. Palgrave Macmillan.

Husain, Z., Dutta, M. and Chowdhary, N., 2014. Is Health Wealth? Results of a Panel Data Analysis. Social Indicators Research, 117(1), 121-143 Springer.

Hausman, J. A., 1978. Specification Test in Econometrics. Econometrica, 46(6) (Nov., 1978), 1251-1271.

Hsiao, C., 2003. Analysis of Panel Data. University of Southern California. Cambridge University Press, 384p.

Hayash, F., 1992. The Estimation of Panel-Data Models with Serial Correlation When Instruments Are Not Strictly Exogenous. Journal of Business Economic Statistics, 10(1), 15-17.

Hiestand, T., 2005. Using pooled Model, Random Model and Fixed Model Multiple Regression to Measure Foreign Direct Investment in Taiwan. International Business and Economics Research Journal, 4(12).

Henderson, C. R., Jr., 1971. The Use of Error Component Models in Combining Cross Section with Time Series Data. Econometrica, 39(2), 397-401.

Karamidehkordi, E., 2010. A Country report: Chalenges Facing Iranian Agriculture and Natural Resource Management in the Twenty First Century. In Human Ecology 389(1), 295-303.

Katchova, A., 2013. Panel Data Models.

Khandker, S. R., Barnes, D. F. and Samad, H. S., 2013. Welfare Impacts of Rural Electrification: A Panel Data Analysis from Vietnam. Economic Development and Cultural Change. 61(3), 659-692. University of Chicago Press.

Liang, Y., Lu, W. and Ying, Z. 2009. Joint Modeling and Analysis of Longitudinal Data with Informative Observation Times. Biometrics, 65(2), 377-384. International Biometric Society.

Li, D., Chen, J. and Gao, J., 2010. Non-parametric time-varying coefficient panel data models with fixed effects. The Econometrics Journal, 14(3), 387-408.

Lee, C. C. and Lee, J., 2010. A Panel Data Analysis of the Demand for Total Energy and Electricity in OECD Countries. The Energy Journal, 31(1), 1-23.

Li, F., 2010. Economic Base Analysis Using Panel Data Regression: A Case Study of the Florida Regional Economy. University of Florida.

Lee, S. H., Eom, T. H. and Xu, H., 2007. Introduction to Panel Data Analysis: Concepts and Practices. Handbook of Research Method in Public Administration, Page: 571.

Mayda, A. M., 2010. International migration: a panel data analysis of the determinants of bilateral flows. Journal of Population Economics, 23(4), 1249-1274. Springer.

Mofya, R. and Hichaambwa, M., 2018. Livlihood Effects of Crop diversification: A Panel Data Analysis of Rural Farm Households in Zambia. In Food security, 10(6), 1449-1462.

Modarres, R. and Sarhadi, A.,2009. Rainfall Trends Analysis of Iran in the Last Half of the Twentieth Century. In Journal of Geophysical research, Vol: 114.

Magnus, J. R. and Muris, C., 2010. Specification of Variance Matrices for Panel Data Models. Econometric Theory, 26(1), 301-310. Cambridge University Press.

Namulondo, R., Bashaasha, B., 2020. Panel Data Analysis of a Crop Diversification Strategy in South-Western and Northern Uganda. The African Economic Research Consortium. Kenya.

Njoroge, L. and Kamau, A. W., 2010. Macroeconomic developments and banks' behavior in Kenya: a panel data analysis. Savings and Development, 34(2), 191-217. Centre for Socio-economic Dynamics and Cooperation of the University of Bergamo.

Neyman, J. and Scott, E., 1998. Consistent estimates based on partially consistent observations. Econometrica. 16(1), 1-32.

Phillips, J. A. and Greenberg, D. F., 2008. A Comparison of Methods for Analyzing Criminological Panel Data. Journal of Quantitative Criminology, 24(1), 51-72.

Pirotte, A. and Madre, J. L., 2011. Car Traffic Elasticities: A Spatial Panel Data Analysis of French Regions. In Journal of Transport Economics and Policy, 45(3), 341-365.

Romero-Jordán, D., del Río, P. and Burguillo, M., 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, 48(2), 315-332. University of Bath.

Soltani, S., Saboohi, R. and Yaghmaei, L., 2012. Rainfall and rainy Days Trend in Iran. In Climate change, 110, 187-213.

Swamy,P. A. V. B. and Arora, S. S.,1972. The Exact Finite Sample Properties of the Estimation of Coefficients in the Error Components Regression Models. Econometrica, 40(2), 261-275.

Searle, R. S., Casella, G. and McCulloch, C. E., 2006. Variance Components. John Wiley & Sons,ING.,publication, New Jeresy, 537p.

Tsounis, N., 2017. Advances in Panel Data Analysis in Applied Economic Research. In International Conference on Applied Economic.

Taofik, A. O., Aribike, E. and Ekum, M.,2018. Panel Predictive Modeling of Agricultural Production Among States in Nigeria. IOSR Journal of Mathematics, 13(5). 76-89. Department of Mathematics and statistics, Ikordo, Lagos State.

Torre, R. and Myrskyla, M., 2011. Income Inequality and Population Health: A Panel Data Analysis on 21 Developed Countries. In Max plank Institute for Demographic Research.

Vergeer, R. and Kleinknecht, A., 2010. The impact of labor market deregulatio on productivity: a panel data analysis of 19 OECD countries (1960— 2004). Journal of Post Keynesian Economics. 33(2). 371-407.

Wooldridge, J. M., 2010 Econometric Analysis of Cross Section and Panel Data. The MIT Press, London, England, 741p.

Agung, I. G. N., 2014. Panel Data Analysis Using Eviews. John Wiley & Sons Ltd, United Kingdom, 541p.

Wallace, T. D. and Hussain, A., 1969. The Use of Error Components Models in Combining Cross Section with Time Series Data. Econometrica. 37(1), 55-72.

White, H., 1980. A Heteroscedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. Econometrica, 48(4), 817-838.

Zulfikar, R., S.Tp, M.M., 2018. Estimation Model and Selection Method of Panel Data Regression: An Overview of Common Effect, Fixed Effect and Random Effect Model. Econometrics. Economy Faculty, University of Islam Kalimantan, MAB Banjarmasin.

CURRICULUM VITAE

Darya Latif RASUL was born in Iraq-Sulaimaniya city in 6-6-1987. She completed her primary, secondary and high school education in Sulaimaniya. She graduated from department of statistics of Sulaimaniya University in 2010. She started her first job as statistician at Sulaimaniya statistics directorate. In 2015 she started her new job at Sulaimaniya Polytechnics University and still continuing. In 2018 she started her MSc. degree study program in the department of statistics of Çukurova University in Turkey-Adana. She is married and she is mother of two children.