Modeling Heteroskedasticity of Crop Yield Distributions: Implications for Normality

Luabeya F. Kapiamba
Ph.D. Student
Departments of Economics & Agricultural Economics
Michigan State University
108 Cook Hall
East Lansing, MI 48824
Tel.(517)-353-7895
Email: kapiamba@msu.edu

Selected Paper Prepared for Presentation at the American Agricultural Economics Association Annual Meeting, Providence, Rhode Island, July 24-27, 2005.

Abstract
The paper analyzes the extent to which ignorance of heteroskedasticity or its inadequate modeling would result in misleading statistical inferences about crop yield distribution. We follow the "detrending mean yield approach" in which we model the conditional mean yield using a panel data model. We assume alternative structures of variance-covariance matrix for the random component. Heteroskedasticity robust and non-robust estimation methods are used before performing a joint normality test on the random component of crop yield data. Our findings provide evidence against the claim that virtually all previous findings of non-normality in crop yields are infected because of the ignorance of heteroskedasticity or its inappropriate modeling. Accounting for heteroskedasticity in crop yield data would matter for validity of evidence against normality only to the extent that its proportion of departure in the data from normal distribution is relatively sizable.

1. Introduction

Over the last decades there has been a considerable empirical work by agricultural economists to determine the appropriate probability distribution model that best characterizes crop yield. The central focus since the early 1970s has been the issue of whether crop yields are normally distributed or not. Modeling crop yield distributions is relevant for many purposes such as (i) estimating crop yield risk, designing and rating crop insurance contracts; (ii) decision-making in agricultural production and risk management under uncertainty; (iii) framing sectoral farm policies.

Modeling crop yield distributions, however, has been quite difficult as substantiated by the number distribution models postulated and investigated in the literature, as well as the resulting empirical work to determine which one best fits the data. Conventional approaches in early studies aimed at estimating yield risk and rating crop insurance contracts used the normal distribution. Following an influential study by Day (1965) that found evidence against normality on crop yield distributions, a view has emerged that crop yields are skewed and do not follow normal distributions. The theoretical basis for this view is built on two related elements: (i) the biological constraints that limit the maximum yield that can be observed and (ii) the environmental factors (e.g.; weather, pest damages) that often affect output. Such elements would make low yields more likely observed (Goodwin and Ker, 1998).

Day's work suggested that distributions of crop yields in the Mississippi cotton, corn and oats have negative skewness (the distribution has a long left tail). Gallagher (1987) reported that soybean yields are negatively skewed; Nelson and Preckel found corn distributions to be negatively skewed given average fertilizer use. Swinton and King (1991), Ramirez (1999), Taylor (1990), Moss and Shonkwiler (1993) also report evidence of negative skewness. In light of these findings, it became clear that failure to recognize this skewed yield distribution leads to underestimating of risk yields with severe

consequences, especially in rating yield risk for crop insurance design purposes. This has led researchers to propose and investigate alternative distribution models.

A relatively recent article by Just and Weninger (1999) suggested that previous findings of skewed yield distributions may be the result of inappropriate detrending and failure to properly model heteroskedasticity. They identify three methodological problems common in yield distributions analysis: (i) misspecification of the nonrandom components of the yield distribution, specifically, the assumption of the linearity in time trend for the mean of the distribution, and the ad hoc modeling of the heteroskedasticity; (ii) misreporting of statistical significance, and (iii) use of aggregate time-series data to represent farm-level yield distributions. They conclude that, one or more of these problems infect virtually all evidence against normality to date.

The modeling of heteroskedasticity in investigating crop yield distributions is the focus of this paper. Heteroskedasticity has long been recognized in statistical analysis of crop yields (Gallagher, 1987); nevertheless it has received less attention and frequently has been handled inadequately in empirical analyses (Yang, Koo and Wilson, 1992). Early studies used the coefficient of variation around the trend to measure the variability in crop production (Hazell, 1984; Weber and Sievers, 1985; Singh and Byerlee, 1990). The underlying assumption was that detrended yields are homoskedastic within the sample period. Following Gallagher (1987), recent studies have tried to account for heteroskedasticity in the nonrandom component of the crop yield; however, different approaches have been adopted on a had hoc basis. There is no common ground on how to model heteroskedasticity nor a consensus on its implications for statistical inference on crop yields distribution.

To which extent would the ignorance of heteroskedasticity or its inadequate modeling result in misleading inferences about the crop yield distribution? This question is the main focus of this paper. Our objective is to evaluate how sensitive is the test for normal distribution to alternative methods for modeling heteroskedasticity in crop yield data. Following Just and Weninger (1999), we test the hypothesis that failure to account for heteroskedasticity or adequately model its structure in crop yield data lead to falsely rejecting normality while it is the appropriate distribution.

Our empirical analysis uses data on Soybeans and Corn grains for 99 counties in Iowa from 19972-2003. The data set is drawn from the National Agricultural Statistics Service (NASS) data base available on the website. The results from the joint test for normal skewness and kurtosis suggest that normal distribution cannot be supported for Corn and Soybeans crop yields even after using alternative estimation methods that are robust in the presence of particular structures of heteroskedasticity. We use the information Matrix (IM) test to investigate this puzzle and found that the proportion of crop yields departure from normal distribution that can be attributable to heteroskedasticity is relatively small, which suggest why correcting for it does not lead to new statistical inferences about the non-normality finding.

The rest of the paper is organized as follows. Section 2 provides a background on the empirical work for testing normality in crop yield distributions and reviews some previous findings. Section 3 presents the models and estimation methods that can used to account for the presence of particular form of heteroskedasticity in the panel data setting. Procedures for testing normality are also presented. In section 4 we describe the data set used, the empirical results and their analysis. We close with a concluding summary in section 5.

2. Background

Estimation of crop yield distribution has been carried out by agricultural economists using two main approaches depending on whether they rely on known parametric distribution or, alternatively, on nonparametric methods. Parametric methods require specifying a functional form and distributional assumptions about the random component (error term). Nonparametric methods are flexible and make no a priori assumption about the error distribution; in this sense, they essentially nest parametric distributions (Goodwin and Ker, 1998).

Under the parametric approach, the typical procedure has been to estimate the conditional mean yield, remove it from data, and study the distribution of the random component; this procedure is referred to as "detrending mean yield approach". The most challenging task in using this approach has been to come up with a well specified functional form for the conditional mean yield. The complexity of economic, behavioral, biophysical and sociological processes makes it difficult to correctly specify the model that represents the crop yield. The typical approach in papers that explored the distributions of crop yield has been to use the deterministic component of yields, which can be adequately represented by a polynomial trend function. The main justification for using deterministic component, as pointed out by Just and Weninger (1999), is that if economic variables move slowly through time (as widely acknowledged in frequency domain literature), then approximation of deterministic component of yields may be sufficient for testing normality regardless of the complexity of underlying process. The assumption here is that the composite of environmental effects on crop yields are captured in the disturbance term of a statistical model.

The review of literature suggests that various algebraic forms of the yield response regression have been used to detrend the data. Likewise, different corrections methods have been applied to account for heteroskedasticity on a had hoc basis. Gallagher (1987) used a frontier model in which the maximum attainable yield, Y^M, is regressed on an annual time trend t; recognizing heteroskedasticity, he used OLS residuals to estimate error variance as a function of a time trend. He then used the standard deviation to create a time-specific index for yield variance, denoted by $VS_{t,}$ and uses it to weight the observations. He substitutes these standardized deviations from this yield frontier as the random variable in the gamma probability density function, from which he estimates the model parameters by maximum likelihood.

Nelson and Preckel (1989) utilized the concept of maximum attainable yield. They represent the deterministic component of yields with an economic variable (fertilizer). Deviations of yield from its maximum were then modeled as a beta distribution conditioned on agricultural inputs. They estimated the two beta parameters by maximum likelihood procedure and claimed that yields in each county are negatively skewed. However, Nelson and Preckel's analysis did not consider heteroskedasticity and correlation of yields among farms.

Moss and Shonkwiler (1993) modeled mean yield of corn as a linear time trend, but allowed the parameters of this trend to be random according to a Kalman Filter. This is referred to as stochastic trend model. They tested the residuals for normality using the Kolmogorov-Smirnov (a nonparametric test) and a parametric test described by Bera and Jarque. They imposed homoskedasticity to maintain the tractability in the Kalman filter. Their findings provided support for negative skewness in U.S. corn yield data for 1930-90. Ramirez (1997) also used an inverse hyperbolic sine transformation in a multivariate non-normal parametric model of yield distributions for U.S. corn Belt corn, soybean, and

wheat from 1950-89. After removing a linear trend, he tested the random component and found heteroscedasticity and nonnormal kurtosis for corn and soybeans.

Goodwin and Ker (1998) modeled the mean yields as an ARIMA process in which they represent the percent deviations of yield from its mean with a nonparametric kernel smother. To account for heteroskedasticity, they considered Goldfel-Quandt parametric and nonparametric heteroskedasticity tests for the ARIMA residuals. In light of their results, they used proportional errors (calculated by dividing each error by its associated error forecast) to model the distributions about the forecasted yields.

Weninger (1999) pointed out, however, that the order in which tests are done invalidates the nonnormality findings. Skewness and kurtosis are tested first under the null hypothesis of homoskedasticity whereas homoskedasticity is subsequently rejected. Their work suggested that previous findings of skewness and nonnormality may be the result of inappropriate detrending and failure to properly model heteroskedasticity.

This literature review suggests that many of the findings on the nonnormal skewness and kurtosis reported in previous studies of crop yield distributions may not be robust to alternative assumptions for modeling heteroskedasticity and estimation methods used to account for it. The next section explores the strategies for modeling and estimating models with particular forms of heteroskedasticity in the context of panel data analysis. A test procedure for testing normality is presented.

3. Empirical Framework

3.1. Modeling Conditional Mean for Crop Yield Data

In testing normality of crop yield distribution in the "detrending" framework it is assumed that yield data can be decomposed into two parts: a deterministic and a random component. The test requires isolating the random component of the yield data. A comprehensive economic model of the deterministic trend could be used for this purpose. However, as indicated above, such model of the conditional mean crop yield is rarely available [see Just and Weninger (1999) for more discussion]. Instead most analyses proceed by assuming that the deterministic trend can be approximated by a low-order trend function. We follow this approach widely used in the literature to keep the results comparable with previous findings.

Estimating conditional mean yield for crop using data from different units of observation over different periods of time can be appropriately handled using methods developed in the context of panel data models. For Soybean and Corn yields observed in 99 counties in Iowa from 1973-2003, the basic unobserved effect model (UEF) can be written as:

$$Y_{it} = X_{it}\beta + \alpha_i + u_{it} \qquad (1)$$

Where Y_{it} is crop yield observed in county i (i = 1, 2,…, N) at period t (t = 1, 2,…,T); α_i are called, unobserved individual effects or unobserved heterogeneity; the α_i are invariant over time, but they are assumed to be different across counties; the u_{it} are called the idiosyncratic errors or idiosyncratic disturbances and they change across time. The X_{it} represent polynomial trend variables. There are different methods for estimating equation (1). We present thereafter the most frequently used of them with particular focus on how heteroskedasticity is modeled and/or corrected.

3.1.1. The Pooled OLS Estimator (POLS)

Under certain assumptions, the pooled OLS estimator can be used to obtain a consistent estimator of β in the model (1). Write the model as

$$Y_{it} = X_{it}\beta + v_{it} \qquad (2)$$

Where $v_{it} = \alpha_i + u_{it}$, $t = 1,\ldots,T$ are the composite errors. For each t, v_{it} is the sum of the unobserved effect and an idiosyncratic error. Assuming that $v_{it} \sim iid(0, \sigma^2)$, that is, for a given X_{it}, there is no serial autocorrelation between observations and, furthermore, errors are not heteroskedastic, a consistent and efficient estimator of β can be obtained by Pooled OLS. However, ignoring the panel structure of the data by assuming that the error terms are iid leads to results that are not appropriate in many cases; due to the presence of α_i in each time period, the composite error will be serially correlated. Even though serial correlation is absent, contemporaneous correlation across panels may still exist. Neverthless, and despite its potential problems, pooled OLS is often used as starting point in applied analyses. Typically, its results are compared to results from models that are better suited for the analysis of panel data.

3.1. 2. Extensions to the Pooled OLS Estimator

Let consider the basic unobserved effects model in (1). As specified above, u_{it} corresponds to the common stochastic error term, and α_i is the individual-specific effect, which is assumed to vary across individuals but is constant over time. The two explicit assumptions about u_{it} are that (i) u_{it} is uncorrelated with $X_{it;}$ and (ii) it varies unsystematicaly across individuals and time. In particular:

- $E(u_{it} \mid X) = 0$ (3)
- $E(u_{it}, u_{js}) \mid X) = 0$ for all $t \neq s$ or $i \neq j$. (4)

In modern econometric parlance (Wooldridge, 2001), we distinguish two basic panel data model depending on whether or not we assume zero correlation between the observed explanatory variables and the unobserved effects.

- Random effects model: α_i is uncorrelated with X_{it} or $Cov(X_{it}, \alpha_i) = 0$;
- Fixed effects model: α_i is correlated with X_{it} or $Cov(X_{it,} \alpha_i) \neq 0$

3.1.2.1. Random Effect Model (RE)

Under the random effects assumptions,

$$Corr(v_{it},v_{is}) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_u^2),\ t \neq s, \quad (5)$$

Where $\sigma_\alpha^2 = Var(\alpha_i)$ and $\sigma_u^2 = Var(u_{it})$. Because the usual pooled OLS standard errors ignore this correlation, they will be incorrect, as will the usual test statistics. We can use generalized least squares (GLS) to estimate model with serial autocorrelation.

Let define $\lambda = 1 - [\sigma_u^2/(\sigma_u^2 + T\sigma_\alpha^2)]^{1/2}$, with $0 < \lambda < 0$. (6)

Then the transformed equation

$$Y_{it} - \lambda\, Y^{*}_{t} = \lambda(X_{it} - X^{*}_{t})'\beta + (v_{it} - \lambda v^{*}_{t}), \quad (7)$$

where the Y^{*} denotes the time average, $1/T\sum Y_{it}$, similarly for X^{*}_{i} and v^{*} . This equation involves a quasi-demeaned data. The GLS estimator is simply the pooled OLS estimator of equation (7). The errors in (7) are serially uncorrelated. When $\lambda = 0$, the random effects estimator is equivalent to pooled OLS.

3.1.2.2. Fixed Effects Model (FE)

The crucial assumption in the fixed effects model is that $cov(X_{it}, \alpha i) \neq 0$. The fixed effects procedure estimates equation (1) by removing the unobserved effect, αi. To see this, consider the basic unobserved effects in equation (1). Now, for each i, average this equation over time. We get

$$Y^{*}_{i} = X^{*}_{i}\beta + \alpha_i + u^{*}_{i} \quad (9)$$

Where $y^{*}_{i} = T^{-1}\sum_{i=1} Y_t$, and so on.

If we substract (9) from (1) for each t, we wind up with

$$Y_{it} - Y^*_i = (X_{it} - X^*_i)'\beta + u_{it} - u^*_{i,}\ t = 1,2,\ldots,T \qquad (10)$$

A pooled OLS estimator that is based on equation (10) is called the fixed effects estimator or the within estimator. A point to notice is that when $\lambda = 1$, the random effect estimator is identical to the fixed effect estimator.

3.1.3. Testing and Correcting for Heteroskedasticity

The presence of heteroskedasticity is of primary interest in this study. Pooled OLS assumes homoskedastic errors for the usual inferences accompanying it to be valid. The FE estimator is based on the assumption that the idiosyncratic errors are homoskedastic within and across panels, and not serially and/or cross-sectional autocorrelated. Likewise, the RE model only assumes serially correlation because of the presence of α_i in the composite error term; it does not account for cross-panels correlation. When these assumptions are violated the FE and the RE estimators and the accompanying inferences procedures are not valid.

In practice, after regression by POLS, there are usual procedures to detect heteroskedasticity, including LM test, Breusch-Pagan's test, White's test, Godfeld and Quandt's test, Bartletts's test, Glejser's test and the like). Once heteroskedasticity is detected, estimation procedures are used to correct for it. However, robustness of estimation is contingent upon the structure of the heteroskedasticity present in the data. This structure is often unknown and assumptions are often made, opening possibilities of misleading inferences. In the context of panel data, while one may allow for heteroskedasticity within panels, we can think of heteroskedasticity across panels, which require a different estimation procedure for making the errors homoskedastic. Further, one may allow cross-sectional correlation to exist, and

within panels, correlation coefficient may be unique for each panel. This can be done by allowing more flexible structure of Ω, the variance-covariance matrix of v_{it}, the composite error term. Possible alternative procedures have been suggested to transform data towards homoskedasticity; some are robust to the structure of heteroskedasticity while others require specifying its form. In this analysis we will use procedures based on Weighted Least Squares (WLS), Feasible Generalized Least Squares (FGLS), and Heteroskedasticity-Robust Standard Errors.

The variance matrix of the disturbance terms can be written as

$$E[uu'] = \Omega = \begin{pmatrix} \sigma_{11}\Omega_{11} & \sigma_{12}\Omega_{12} & . & \sigma_{1n}\Omega_{1n} \\ \sigma_{21}\Omega_{12} & \sigma_{22}\Omega_{22} & . & \sigma_{2n}\Omega_{2n} \\ . & . & . & . \\ \sigma_{n1}\Omega_{n1} & \sigma_{n2}\Omega_{n2} & . & \sigma_{nn}\Omega_{nn} \end{pmatrix}. \tag{10}$$

In order for the Ω_{ij} matrices to be parametrized to model cross section correlation, they must be square (balanced panels).

For the classic POLS regression model, we have

$$\begin{aligned} &E[v_{it}] = 0 \\ &Var[v_{it}] = \sigma^2 \\ &Cov[v_{it}, v_{js}] = 0 \text{ if } t \neq s \text{ or } i \neq j. \end{aligned} \tag{11}$$

This amount to assuming that Ω has the structure given by

$$\Omega = \begin{pmatrix} \sigma^2 I & 0 & . & 0 \\ 0 & \sigma^2 I & . & 0 \\ . & . & . & . \\ 0 & 0 & . & \sigma^2 I \end{pmatrix}. \tag{12}$$

In many cross-sectional datasets, the variance for each of the panels will differ. The heteroskedasticity model is specified by including the panels heteroskedasticity structure, which assumes that

$$\Omega = \begin{pmatrix} \sigma_1 I & 0 & . & 0 \\ 0 & \sigma_2 I & . & 0 \\ . & . & . & . \\ 0 & 0 & . & \sigma_n I \end{pmatrix}. \tag{13}$$

We may wish to assume that the error terms of panels are correlated, in addition to having different scale variances (heteroskedasticity across panels). This is achieved by specifying more general structure of the variance-covariance matrix as

$$\Omega = \begin{pmatrix} \sigma_{11} I_{11} & \sigma_{12} I_{12} & . & \sigma_{1n} I_{1n} \\ \sigma_{21} I_{12} & \sigma_{22} I_{22} & . & \sigma_{2n} I \\ . & . & . & . \\ \sigma_{n1} I_{n1} & \sigma_{n2} I_{n2} & . & \sigma_{nn} I_{nn} \end{pmatrix}. \tag{14}$$

We may further consider more general structure of Ω in order to allow for autocorrelation within panels. To do this the individual identity matrices along the diagonal of Ω may be replaced. Among the possible specifications, we may assume a structure with (i) no autocorrelation; (ii) serial autocorrelation where the correlation parameter is common for all panels; (iii) serial correlation where the correlation parameter is unique for each panel [see Stata Cross-Sectional Time Series Reference manual Release 8]. Estimation procedures such as the Feasible Generalized Least Squares (using either the estimated cross-section residual variances or the cross-section residual covariance matrix as weights), White Heteroskedasticity covariance can be used by specifying the appropriate structure of the variance matrix assumed.

3.2. Testing for Crop Yield Normality

In conducting normality tests we use the omnibus or joint moment test strongly recommended by Just and Weninger (1999), such the Skewness and kurtosis tests. There are several ways of measuring skewness and kurtosis but the most well known are Pearson's (1905) skewness and kurtosis. Many tests have been defined using Pearson's skewness and kurtosis statatistics. In this paper, we use the omnibus test (K^2 test) recommended by D'Agostino, Belanger, and D'Agostino, Jr (DBD). We refer to DBD's skewness and Kurtosis statistics as $Z(b_1)^{1/2}$ and $Z(b_2)$ respectively. Under the null hypothesis of normal residuals, both $Z(b_1)^{1/2}$ and $Z(b_2)$ are distributed as approximately normal (0,1). DBD's omnibus K^2 statistic is constructed as

$$K^2 = [Z(b_1)^{1/2}]^2 + [Z(b_2)]^2 \qquad (15)$$

and is approximately distributed as chi-squared (χ^2) with two degrees of freedom. The sketest command in STATA implements the test described by DBD (1990) with the empirical correction developed by Royston (1991).

White (1980)'s Information Matrix (IM) procedure developed by Cameron and Trevedi (1990) offers another appealing way to perform this kind of omnibus test. The IM procedure exploits the well-known property that, at the model, the sum of the Hessian of the log-likelihood and the outer product of the score has zero expectation. It tests the hypothesis that the information matrix equality holds, that is the hypothesis that:

$$H0: E\{D^2 \log f(X, \theta_0)\} = E\{ D \log f(X,\theta_o).D'\log f(X,\theta_0). \qquad (15)$$

Where $\log f(X,\theta_0)$ is the log-likelihood for the random variable X, θ_o is the probability limit of the associated (quasi) maximum likelihood estimator, and the (D)log f and (D^2)log f are the gradient(score) vector and the Hessian matrix of the log likelihood. In the normal model, the IM test using the MLE is

the Jarque-Bera (1980) test for skewness and nonnormal kurtosis. If the H_0 is rejected, the IM test provides an orthogonal decomposition of the source of the deviation from the IM of a normal distribution into part due to heteroskedasticity, skewness and kurtosis.

4. Empirical Estimation and Results

4.1. Data Description

Empirical estimation and test for normality are carried out using county-level data on corn grain and soybeans yield using cross-sectional time-series from 99 counties in Iowa, which is the main region for the production of these two crops in the United States. To avoid selection bias problem we have included in our sample all the counties from the 9 crop reporting districts. The data used for this empirical estimation cover the period from 1972-2003. The choice of starting point was based on the fact that it was only at the beginning of the seventies that the hybrid corn was widely adopted in the cropping system in Iowa. This, we think, could help control for variability due to different patterns in varieties adoption across counties in the State.

The data set is drawn from the National Agricultural Statistics Service (NASS) website, which provides valuable time series data for different crops by state, district, and county. These are aggregate time-series data and, unfortunately, they cannot be used to reflect farm-level randomness which is the most relevant in Multiple Peril Crop Insurance Programs. Nevertheless, the growing interest in Area-yield crop insurance in the recent years is increasing the need for county-level analysis of crop yield distribution

4.2. Analysis of Results

Following the detreding approach to analyzing crop yield distributions, the deterministic component of yield was estimated by fitting a polynomial trend to each county's yield series, where the polynomial degree is determined by the data. Based on F-tests with 5 % significance, a polynomial degree of order greater than 2 was found inappropriate for the data. The results we present here are obtained from a polynomial degree of order 1 and 2 for both Corn for grains and Soybeans.

The basic unobserved effects model (UEM) in equation (1) was estimated using alternative regression methods presented in section 3.1. Pooled OLS estimation was used as starting point, and its results are compared to results from regressions that are better suited for the analysis of panel data when heteroskedasticity, and eventually autocorrelation are modeled. Results are presented in tables 1-4.

4.2.1. Results for heteroskedasticity

Results for heteroskedasticity are obtained by performing the appropriate tests after POLS. Tests were used for both multiplicative and unrestricted heteroskedasticity.

Table1. Breush_Pagan / Cook-Weisberg (BP/CW) Test for Heteroskedasticity

Iowa (1972-2003): Corn and Soybeans County-level yields

Crop	regression	polynomial degree	Chi2(1)	Prob. > Chi2(1)
Corn	Pooled OLS	1	1.66	0.197
Corn	Pooled OLS	2	22.56	0.00
Corn	Fixed Effects	1	1.66	0.197
Corn	Fixed Effects	2	22.56	0.00
Soybeans	Pooled OLS	1	194.04	0.00
Soybeans	Pooled OLS	2	164.26	0.00
Soybeans	Fixed Effects	1	194.04	0.00
Soybeans	Fixed Effects	2	157.86	0.00

Heteroskedasticity is investigated using Breusch-Pagan / Cook-Weisberg (BP/CW) test for assumption of linear regression model that the residuals are homoskedastic, ie., have constant variance against

multiplicative heteroskedasticity. In all cases, except for corn yield first-order polynomial trend, the test provides strong evidence that error components of crop yield are heteroskedastic. The BP/CW test is powerful as it is restrictive in regard to the structure of the heteroskedasticity. The results based on the White test for unrestricted heteroskedasticity [see table 4] reject the null of constant variance in all cases.

4.2.2. Tests for Normality

After detecting heteroskedasticity, we use alternative estimation procedures to account for it. By detrending approach we isolate the random component on which we perform the normality tests. The results are summarized as follows for different yields and polynomial order fitted. Table 2 presents the test results for skewness, normal kurtosis, and the joint test for normality on Corn yield distribution. The p-value are computed for the Chi2-statistics. The results suggest that, regardless of polynomial trend fitted and the estimation techniques used, the null hypothesis of zero skewness and normal kurtosis is rejected with p-value of almost zero. The Chi2 for the joint test of normality rejects normality distribution for Corn Yield. For the linear trend the Chi2 only change marginally from 540.32 in the pooled OLS residuals to 535.78 in the model accounting for heteroskedasticity and correlation across panels. This change is of 470.0 to 393.09 in the quadratic trend specification. In any case, the normality distribution is rejected with the same p-value.

***Table2. Normality Tests: County-level Corn Yields (Iowa: 1972-2003*)**

Polynomial Order Fitted: 1

Regression	**Pr(Skewness)**	**Pr(Kurtosis)**	**Chi2(2)**	**Prob>Chi2(2)**
1. Pooled OLS	0.00	0.00	540.32	0.00
2. Variance-weighted Pooled OLS	0.00	0.00	510.70	0.00
3. Robust standard errors	0.00	0.00	540.32	0.00
4. Cochrance-Orcutt AR(1),semi- robust stand.	0.00	0.00	528.09	0.00
5. Generalized Least Squares	0.00	0.00	540.32	0.00
6. Fixed Effect	0.00	0.00	540.32	0.00
7. Random Effect	0.00	0.00	540.32	0.00
8. Homoskedastic with panel-specific AR(1)	0.00	0.00	539.94	0.00
9. Heteroskedastic with panel-specific AR(1)	0.00	0.00	538.72	0.00
10. Heteroskedasticity & correl. across panels	0.00	0.00	535.78	0.00

Polynomial Order Fitted: 2

Regression	**Pr(Skewness)**	**Pr(Kurtosis)**	**Chi2(2)**	**Prob>Chi2(2)**
1. Pooled OLS	0.00	0.00	470.00	0.00
2. Variance-weighted Pooled OLS	0.00	0.00	499.62	0.00
3. Robust standard errors	0.00	0.00	465.80	0.00
4. Cochrance-Orcutt AR(1), semi- robust stand.	0.00	0.00	466.88	0.00
5. Generalized Least Squares	0.00	0.00	470.11	0.00
6. Fixed Effect	0.00	0.00	470.11	0.00
7. Random Effect	0.00	0.00	470.11	0.00
8. Homoskedastic with panel-specific AR(1).	0.00	0.00	471.50	0.00
9. Heteroskedastic with panel-specific AR(1).	0.00	0.00	475.21	0.00
10. Heteroskedastic & correlation. across panels	0.00	0.00	393.09	0.00

The results in table 3 summarize the test statistics for normality test on county-level Soybeans. As for Corn, the random component of soybeans yield is found to have skewness, nonnormal kurtosis, and is not normally distributed. No substantial change in the Chi2-statistics is noticeable from one regression to the other.

Table 3. Normality Tests : Country-level Soybeans Yields (Iowa:1972-2003)

Polynomial Order Specification:1	Skewness and Kurtosis Test of Normality			
Regression	**Pr(Skewness)**	**Pr(Kurtosis)**	**Chi2(2)**	**Prob>Chi2(2)**
1. Pooled OLS	0.00	0.00	267.89	0.00
2. Variance-weighted Pooled OLS	0.00	0.00	272.18	0.00
3. Robust standard errors	0.00	0.00	267.89	0.00
4. Cochrance-Orcutt AR(1), semi- robust stand.	0.00	0.00	271.22	0.00
5. Generalized Least Squares	0.00	0.00	267.89	0.00
6. Fixed Effect	0.00	0.00	267.89	0.00
7. Random Effect	0.00	0.00	267.89	0.00
8. Homoskedastic with panel-specific AR(1).	0.00	0.00	252.68	0.00
9. Heteroskedastic with panel-specific AR(1).	0.00	0.00	246.68	0.00
10. Heteroskedastic & correlation. across panels	0.00	0.00	246.12	0.00

Polynomial Order Specification:2				
Regression	**Pr(Skewness)**	**Pr(Kurtosis)**	**Chi2(2)**	**Prob>Chi2(2)**
1. Pooled OLS	0.00	0.00	268.29	0.00
2. Variance-weighted Pooled OLS	0.00	0.00	251.76	0.00
3. Robust standard errors	0.00	0.00	268.29	0.00
4. Cochrance-Orcutt AR(1), semi- robust stand.	0.00	0.00	282.10	0.00
5. Generalized Least Squares	0.00	0.00	268.29	0.00
6. Fixed Effect	0.00	0.00	268.29	0.00
7. Random Effect	0.00	0.00	268.29	0.00
8. Homoskedastic with panel-specific AR(1).	0.00	0.00	254.78	0.00
9. Heteroskedastic with panel-specific AR(1).	0.00	0.00	249.61	0.00
10. Heteroskedastic & correlation. across panels	0.00	0.00	244.04	0.00

The results presented here are puzzling; why is the normality tests performed on crop yield's error component from regression accounting for the presence for heteroskedasticity lead to the same inference at same the p-value, with only marginal change in the test statistic? To investigate this question we perform the information matrix test and its orthogonal decomposition into test for heteroskedasticity, skewness and kurtosis due to Cameron and Trivedi (1990). Table 4 presents the results of the test performed on the pooled OLS residuals.

Table 4. Test for Unrestricted Heteroskedasticity and Information Matrix Test Decomposition

	Corn (1)			Corn (2)			Soybeans (1)			Soybeans (2)		
Source	Chi2	DF	p-value	Chi2	DF	p-value	Chi2	DF	p-value	Chi2	DF	p-value
Heteroskedasticity	56.9	2	0.00	94.4	2	0.00	151.0	2	0.00	245.1	4	0.00
Skewness	237.7	1	0.00	290.2	2	0.00	223.2	1	0.00	195.8	2	0.00
Kurtosis	60.57	1	0.00	52.6	1	0.00	26.63	1	0.00	24.5	1	0.00
TOTAL	355.2	4	0.00	437.2	7	0.00	400.8	4	0.00	465.4	7	0.00

The (.) indicates the degree of polynomial trend fitted for each crop. Based on White Test for H0: Homoskedasticity against Ha: unrestricted heteroskedasticity, he results suggest that there is strong evidence of heteroskedasticity regardless of the polynomial trend model fitted. The decomposition of the information matrix test provides a chi2 statistic to test the hypothesis that the skewness parameter is zero against the alternative that it is different from zero. The Chi2 for testing the normal kurtosis hypothesis is also computed and the p-values provided. The results support the evidence that Corn for grains and Soybeans are skewed and are characterized by non-normal kurtosis.

The decomposition of the departure from normal distribution provides interesting insights on why correcting for heteroskedasticity led to the same inference on normality test as in the model estimated with heteroskedasticity error component. The results in table 4 suggest that heteroskedasticity accounts only for 16 and 21.6 % of the Corn yields departure from the normal distribution in the model fitted on linear trend and quadratic trend respectively. The rest of departure is attributable to the presence of skewness and nonnormal kurtosis, with the former accounting for roughly 67 %. The same can be said for Soybean, though heteroskedasticity seems to be relatively severe, accounting for more than a third of

the departure from normal distributions in linear trend model (37 %) and quadratic trend estimation (52 %). The low proportion of heteroskedasticity among the sources of crop yield data's departure from normal distribution explains why the distribution displays nonnormal shape even after correcting for heteroskedasticity by appropriate estimation methods. In other words, the relative importance of skewness is so high that the distribution remains skewed even after removing heteroskedasticity.

5. Summary and Concluding Remarks

This paper investigated the conflicting evidence about the distribution of crop yield. We focused on one major potential problem that arises from ignoring or inadequately modeling heteroskedasticity in the crop yield data when using detrending approach to testing normality. We adopted unobserved effects model for modeling conditional crop yield, assuming that the deterministic component of yields can be approximated by a smooth function of time, and proposed methods procedures to account for heteroskedasticity. The empirical work was performed using the data on Corn for grains and Soybean yields on a sample of 99 counties from the state of Iowa for the period from 1972-2003.

Using both Breusch-Pagan and White test we found evidence of heteroskedasticity in crop yield data. Different regressions were then performed using alternative methods to correct for heteroskedasticity prior to isolating the random component of crop yield on which the normality test is performed. The omnibus test for normality on the pooled OLS residuals as well as the residuals from well-suited methods in presence of heteroskedasticity led to the rejection of the hypothesis that the crop yields are normally distributed. The Test statistics only display marginal change from pooled OLS to other robust estimation methods, suggesting that findings for nonnormal crop yield distribution are robust to the presence of particular structure of heteroskedasticity in the data.

To investigate these puzzling results, we use the Cameron and Trivedi decomposition of the information matrix test. The results suggested that heteroskedasticity, though present in the data, accounts only for a relatively small part of the crop yield data's departure from the normal distribution. Skewness explains more than half of the departure making the distribution nonnormal even after correcting for heteroskedasticity.

The results of this paper, though preliminary, provide support against the claim that virtually all previous findings on non-normality of crop yield distributions may be infected by the ignorance and/or the inadequate modeling of heteroskedasticity. When heteroskedasticity is detected, ignoring it or inadequately modeling it does not necessarily result in misleading inference about the distribution of crop yields. Its relative importance in the data is key in determining the extent to which it matters for the validity of inference about the distribution of crop yield.

References

1. Atwood, J., S. Shaik and M. Watts. “Can Normality of Yields Be Assumed for Crop Insurance?” Canadian Journal of Agricultural Economics 50(2002):171-84.

2. D'Agostino, RB, Belanger AJ, D'Agostino RB Jr. Tests of Normality and Other Goodness-of-Fit tests. Presented jointly with other authors. Contributed paper at the Northeast SAS User's Group, Fourth Annual Conference, Greenwich, Connecticut, November 1991.

3. Day, R.H. “Probability Distributions of Field Crop Yields.” Journal of Farm Economics 47(1965):713-41.

4. Epplin F. M. “Wheat Yield Response to Changes in Production Practices Induced by Program Provisions.” Journal of Agricultural and Resource Economics 22(1997): 333-44.

5. Gallagher, P. “U.S. Soybean Yields: Estimation and Forecasting with Nonsymmetric Disturbances.” American Journal of Agricultural Economics 71(1987):796-803.

6. Goodwin, B.K. and A.P. Ker. “Nonparametric Estimation of Crop Yield Distributions: Implications for Rating Group Risk (GRP) Crop Insurance Contracts.” American Journal of Agricultural Economics 80(1998): 139-53.

7. Greene, W.H. Econometric Analysis, 5th ed., New York: Prentice-Hall, 2003.

8. Just, R.E. and Q. Weninger. “ Are Crop Yields Normally Distributed?” American Journal of Agricultural Economics 81(1999):287-304.

9. Ker, A.P. and K. Cobble. “Modeling Conditional Yield Densities.” American Journal of Agricultural Economics 85(2003): 291-304

10. Moss, C.B. and J.S. Shonkwiler. “Estimating Yield Distributions with a Stochastic Trend and Nonnormal Errors.” American Journal of Agricultural Economics 75(1993):1056-62.

11. Mukherjee, C., H. White., and M. Wuyts. Econometrics and Data Analysis for Developing Countries. Routledge, New York, 1998.

12. Nelson, C.H. "The Influence of Distribution Assumptions on the Calculations of Crop Insurance Premia." North Central Journal of Agricultural Economics 12(1990):71-8.

13. Nelson, C.H. and P.V. Preckel " The Conditional Beta Distribution as a Stochastic Production Function." American Journal of Agricultural Economics 71(1989):370-78.

14. Norwood, B., M.C. Roberts. and J.L. Lusk. "Ranking Crop Yield Models Using Out-of-Sample Likelihood Functions. "American Journal of Agricultural Economics 86(2004): 1032-1043.

15. Ramirez, O.A. "Estimation and Use of Multivariate Parametric Model for Simulating Heteroscedastic, Correlated, Non-normal Random Variables: The Case of Corn Belt Corn, Soybean and Wheat Yields." American Journal of Agricultural Economics 79 (1997): 291-305.

16. 16.Ramirez, O.A., S. Misra, . and J. Field . " Crop Yield Distributions Revisited." American Journal of Agricultural Economics 85(2003):108-20.

17. Rawlings, J.O. Applied Regression Analysis: A Research Tool, Pacific Grove, CA: Woodsworth & Brooks/Cole, 1988.

18. Swinton, S. and R.P. King. "Evaluating Robust Regression Techniques for Detrending Crop Yield Data with Non-normal Errors." American Journal of Agricultural Economics 73(1991):446-61.

19. Taylor, C.R. " Two Practical Procedures for Estimating Multivariate Nonnormal Probability Density Functions." American Journal of Agricultural Economics 72(1990): 210-17.

20. Wooldridge, J.M. Econometric Analysis of Cross Section and Panel Data. The MIT Press, Cambridge, 2003.

21. Yang, S., W.W. Koo and W.W. Wilson. "Heteroskedasticity in Crop Yield Models". Journal of Agricultural and Resource Economics 17(1992): 103-9.