

ECONOMETRICS PROJECT

OLS REGRESSION MODEL

FACTORS AFFECTING TOTAL FOODGRAIN

PRODUCTION IN INDIA

(1950-2016)

NAME – Lakshay Kumar

ROLL NO – ECO/18/110

CONTENT

1. Introduction

2. The Theoretical Background

Variables used – Dependent
Independent

3. Econometric Methodology

Method of Ordinary least squares (OLS)

4. Regression

Initial Regression Model
Joint Significance Test

5. Tests for violations of OLS Assumptions

Multicollinearity
Heteroscedasticity
Autocorrelation
Testing Normality of Residuals

6. Empirical Results

Graphs
Relationship Graphs
Final OLS Interpretation

7. Conclusion

Introduction

Foodgrains are small, hard, dry seeds, with or without attached hulls or fruit layers, harvested for human or animal consumption.

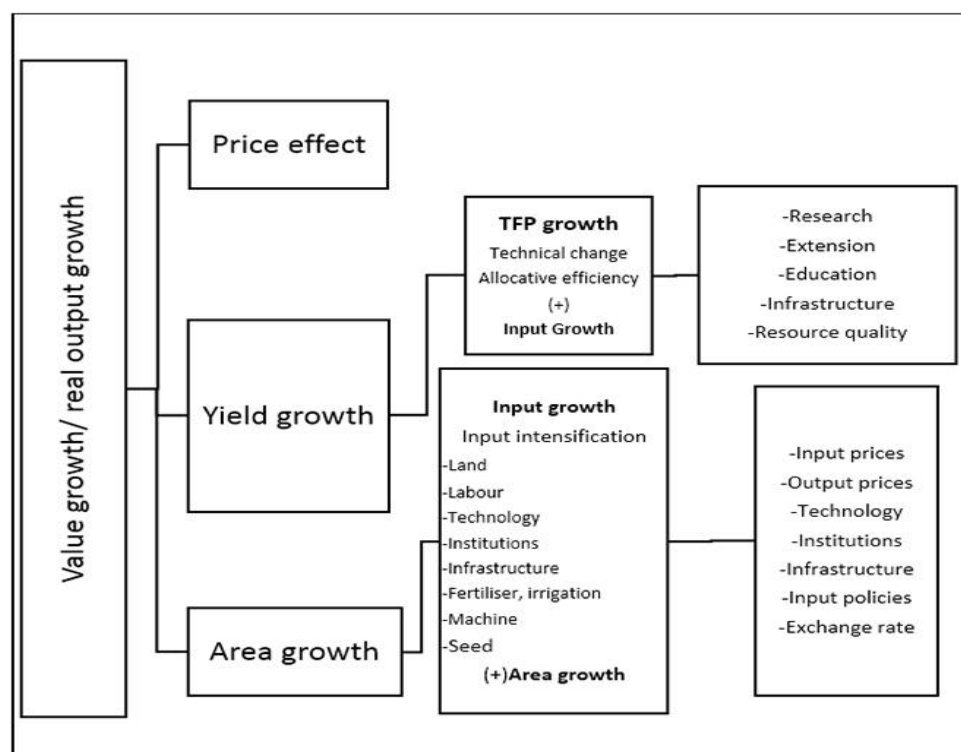
At present, India's rates of food grain availability are worrisome: at 487 grams per person per day. (It was 468.7 grams per person per day in 1961.)

In a report by NITI Aayog (2018) on supply and demand projections for the agriculture and allied sector, the demand for food grains in India is expected to increase by seven percent from 255 million tonnes in 2016-17 to 272 million tonnes by 2020-21.

It is in this context that it becomes important to examine the major factors that affect productivity of food grains in the country. Some of these food grains are also an important component in ensuring availability of high-value products—such as milk and meat—as animal feed.

This paper offers an analysis about the extent of impact of - area used for production of foodgrains, production of rice and wheat on total foodgrain productivity in the nation.

The Theoretical Background



Variables used –

Total Production of Foodgrains in India (Dependent Variable) showcases the production output of foodgrains in the country for that particular year (in million tonnes)

Total area utilized for Production (1st Explanatory Variable) showcases the total area in the country being utilized for production of foodgrains in that particular year (in million hectares)

Yield of Foodgrains (2nd Explanatory Variable) depicts yield of foodgrains in kg per hectare of land under production of foodgrains.

Literature Review –

1. Manojit Chattopadhyay and Subrata Kumar Mitra published their paper on Comparative Decision models for anticipating shortage of food grain production in India in 2016. This paper attempts to predict food shortages in advance from the analysis of rainfall during the monsoon months along with other inputs used for crop production, such as land used for cereal production, percentage of area covered under irrigation and fertiliser use.
2. Advance information of food shortage can help policy makers to take remedial measures in order to prevent devastating consequences arising out of food non-availability. Their paper showed positive relation between area used for production and production of major crops grown with total foodgrains production.

Null Hypothesis – above factors do not affect production of foodgrains

Alternate Hypothesis - above factors affect production of foodgrains

Econometric Methodology

Method of Ordinary least squares (OLS)

Ordinary Least Squares (OLS) method is widely used to estimate the parameter of a linear regression model. OLS estimators minimize the sum of the squared errors (a difference between observed values and predicted values). While OLS is computationally feasible and can be easily used while doing any econometrics test, it is important to know the underlying assumptions of OLS regression. This is because lack of knowledge of OLS assumptions would result in its misuse and give incorrect results for the econometrics test completed. The importance of OLS assumptions cannot be overemphasized. Following are the assumptions of OLS Method:

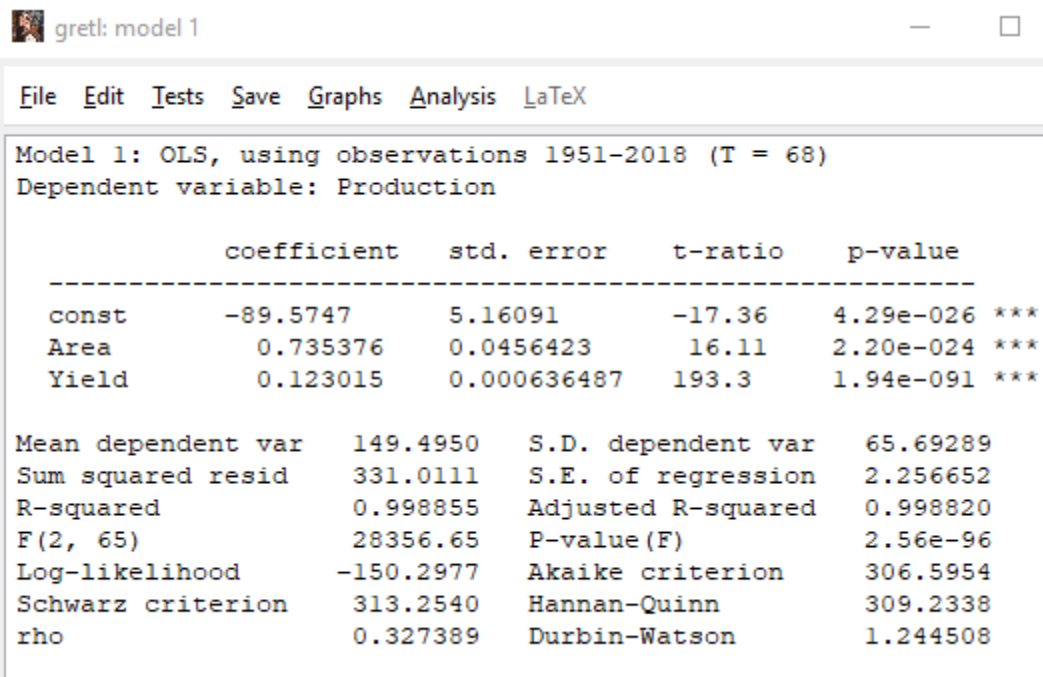
- 1) The linear regression model is “linear in parameters” and correctly specified.
- 2) The Values of Explanatory variables are stochastic.
- 3) Given Values of Explanatory variables mean of the error term is 0.
- 4) Existence of homoscedasticity.
- 5) There is no multi-co linearity (or perfect co linearity)
- 6) Number of Observations should be more than number of explanatory variables.
- 7) Error terms should be normally distributed.

Regression

Initial Regression Model

$$\text{ProdF} = \beta_1 + \beta_2 \text{Area} + \beta_3 \text{Field} + U_i$$

After regressing this model (Through OLS) in gretl following results were obtained



The screenshot shows the 'gretl: model 1' window. The title bar includes a small icon and the text 'gretl: model 1'. Below the title bar is a menu bar with 'File', 'Edit', 'Tests', 'Save', 'Graphs', 'Analysis', and 'LaTeX'. The main content area displays the following text:

Model 1: OLS, using observations 1951-2018 (T = 68)
Dependent variable: Production

	coefficient	std. error	t-ratio	p-value	
const	-89.5747	5.16091	-17.36	4.29e-026	***
Area	0.735376	0.0456423	16.11	2.20e-024	***
Yield	0.123015	0.000636487	193.3	1.94e-091	***

Mean dependent var	149.4950	S.D. dependent var	65.69289
Sum squared resid	331.0111	S.E. of regression	2.256652
R-squared	0.998855	Adjusted R-squared	0.998820
F(2, 65)	28356.65	P-value (F)	2.56e-96
Log-likelihood	-150.2977	Akaike criterion	306.5954
Schwarz criterion	313.2540	Hannan-Quinn	309.2338
rho	0.327389	Durbin-Watson	1.244508

From this we can make out our model as

$$\text{Production} = -89.57 + 0.735 \text{Area} + 0.123 \text{Yield}$$

From the above results we can conclude following things:

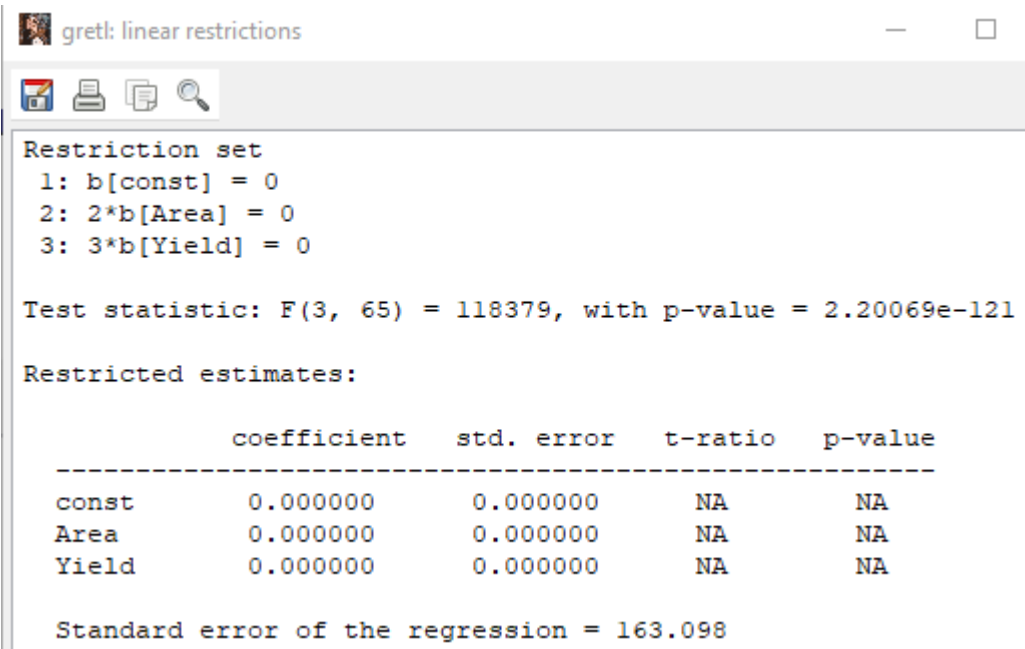
- 1) t ratio's are significant influence for every dependent variable ($|t| > 2$) showing that all the explanatory variables have influence on the Dependent variable i.e. Area, Yield have influence on Production

- 2) Also, the Signs of the coefficients also satisfy our economic theory that states that Area, and Yield have positive relation with Production.
- 3) R squared values is high with low t ratio's which is the 1st sign of multicollinearity which violates our OLS assumptions.

Joint Significance Test

H_0 : Area under production of foodgrains, Yield (in kg) from per hectare land do not affect Production of foodgrains

H_A : Not H_0



```

gretl: linear restrictions
-----
Restriction set
1: b[const] = 0
2: 2*b[Area] = 0
3: 3*b[Yield] = 0

Test statistic: F(3, 65) = 118379, with p-value = 2.20069e-121

Restricted estimates:

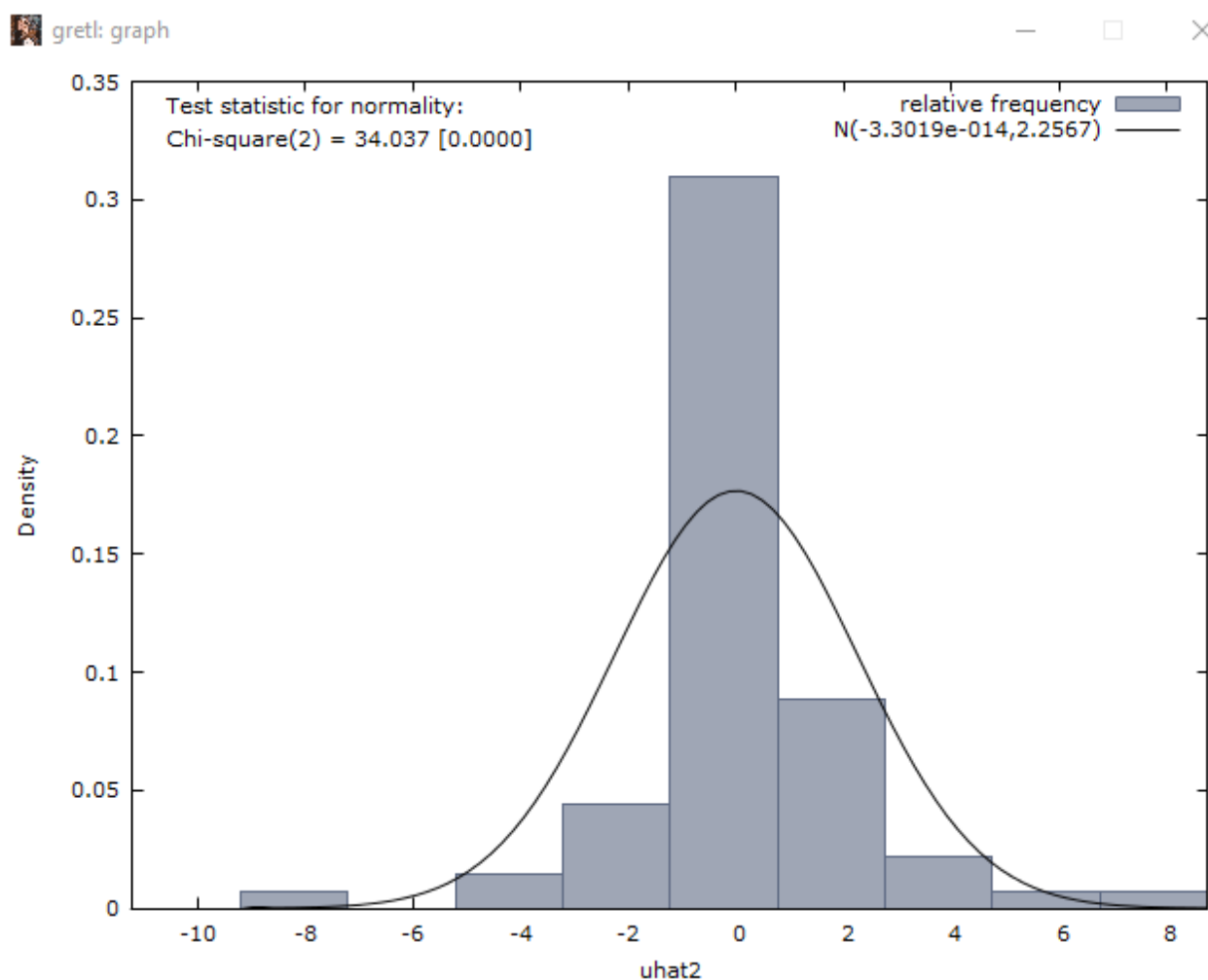
      coefficient    std. error    t-ratio    p-value
-----
const      0.000000      0.000000      NA        NA
Area       0.000000      0.000000      NA        NA
Yield      0.000000      0.000000      NA        NA

Standard error of the regression = 163.098
  
```

Here our null hypothesis was that Yield, Area together have no joint significance on the Production but as we got P value 2.20069e-121 (less than level of significance i.e. 0.05) hence we reject null; so Yield, Area together affect ProdF.

Testing Normality of Residuals (On Original Model with Autocorrelation)

Normality of the error distribution-If the error terms are not normally distributed then the forecasts, confidence intervals, yielded by a regression model may not be BLUE (BEST LINEAR UNBIASED ESTIMATOR)



H₀: Residuals are normally distributed

H_A: Residuals are not normally distributed

On the basis of p value, data is not normally distributed.

Tests for violations of OLS Assumptions

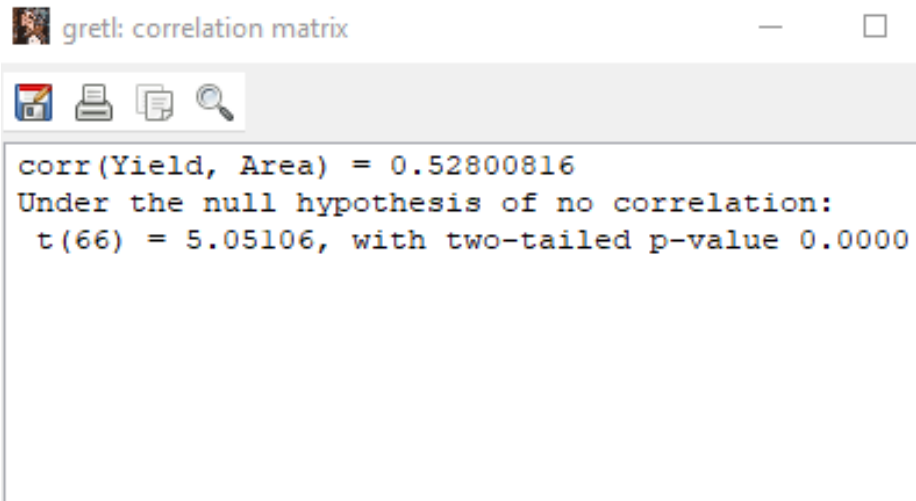
- 1) **Multicollinearity**-Multicollinearity is the occurrence of high intercorrelations among independent variables in a multiple regression model. It can lead to skewed or misleading results. This correlation is a problem because independent variables should be independent.

Testing Multicollinearity

H₀: No Multicollinearity

H_A: Multicollinearity is present

A. Correlation Coefficient Matrix



```
gretl: correlation matrix  
corr(Yield, Area) = 0.52800816  
Under the null hypothesis of no correlation:  
t(66) = 5.05106, with two-tailed p-value 0.0000
```

In this any coefficient (mod value) greater than 0.8 represents a case of high collinearity.

From above results show that Yield does not have collinearity with Area.

B. Variance Inflation Factor

```
gretl: collinearity

Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

      Area      1.387
      Yield      1.387

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation coefficient
between variable j and the other independent variables

Belsley-Kuh-Welsch collinearity diagnostics:

variance proportions

lambda   cond   const   Area   Yield
2.904    1.000   0.000   0.000   0.012
0.094    5.546   0.006   0.003   0.757
0.001   47.062   0.994   0.996   0.231

lambda = eigenvalues of inverse covariance matrix (smallest is 0.00131126)
cond    = condition index
note: variance proportions columns sum to 1.0
```

VIF values less than 10 indicate no collinearity problem. Though this is not the apex test of Multicollinearity, so we move to auxiliary regressions. (But since standard errors are very low, so multicollinearity will not be a problem)

C. Auxiliary Regression

Auxiliary Regression with Area as dependent (Model 3 window)

```
gretl: model 3

File Edit Tests Save Graphs Analysis LaTeX

Model 3: OLS, using observations 1951-2018 (T = 68)
Dependent variable: Area

      coefficient      std. error      t-ratio      p-value
-----
const      111.985         1.92585        58.15        2.01e-058 ***
Yield       0.00736314        0.00145774      5.051        3.70e-06 ***

Mean dependent var      120.9704      S.D. dependent var      7.112625
Sum squared resid      2444.527      S.E. of regression      6.085909
R-squared                0.278793      Adjusted R-squared      0.267865
F(1, 66)                25.51321      P-value(F)              3.70e-06
Log-likelihood           -218.2792      Akaike criterion        440.5584
Schwarz criterion        444.9974      Hannan-Quinn            442.3173
rho                     0.770616      Durbin-Watson            0.318791
```

Following the 'Klein Rule of Thumb' as R-squared value in model 3 window is smaller than overall R-squared value (model 1 window) so the multicollinearity may not be a problem.

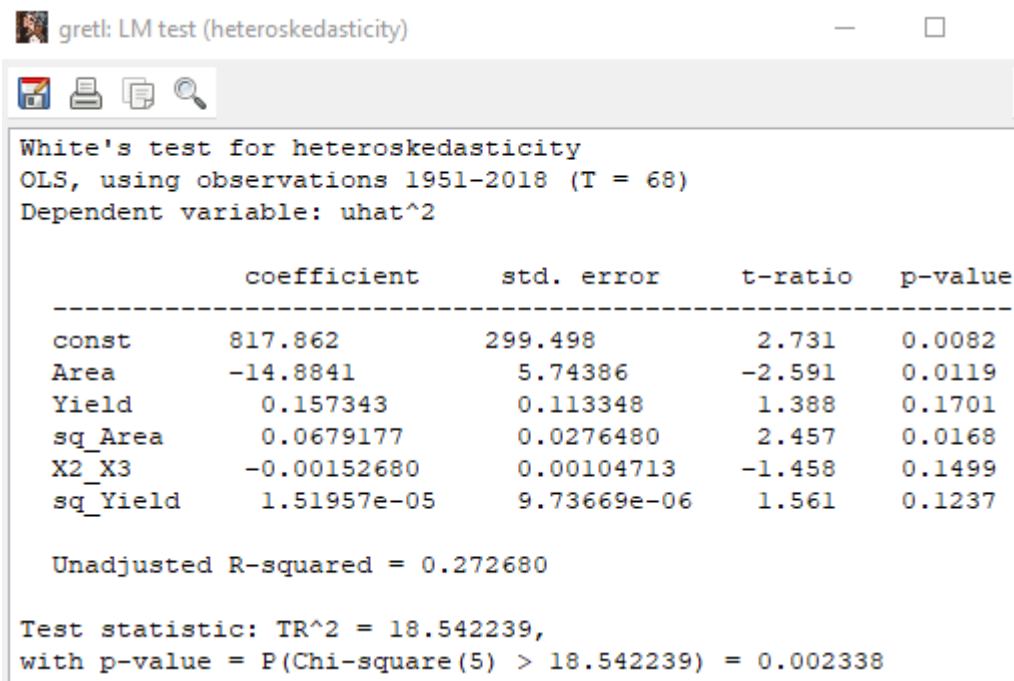
- 2) **Heteroskedasticity** -Let's Recall that OLS makes the assumption that the variance of the error term is constant (Homoscedasticity). If the error terms do not have constant variance, they are said to be heteroscedastic. The existence of heteroscedasticity is a major concern in the application of regression analysis, including the analysis of variance, as it can invalidate statistical tests of significance.

Testing Heteroscedasticity

H₀: Homoskedasticity

H_A: Heteroskedasticity

A. White's Test



Since p-value = $P(\text{Chi-square}(5) > 18.542239) = 0.002338$ is a lot smaller than 0.05 (general level of significance) so we reject null hypothesis, so heteroskedasticity is present in our data.

Remedy for Heteroskedasticity –



gretl: model 2

— □

File Edit Tests Save Graphs Analysis LaTeX

Model 2: WLS, using observations 1951–2018 (T = 68)

Dependent variable: Production

Variable used as weight: Yield

	coefficient	std. error	t-ratio	p-value	
const	-105.504	7.02891	-15.01	8.40e-023	***
Area	0.865659	0.0603987	14.33	8.51e-022	***
Yield	0.122992	0.000670649	183.4	5.85e-090	***

Statistics based on the weighted data:

Sum squared resid	484398.6	S.E. of regression	86.32663
R-squared	0.998567	Adjusted R-squared	0.998523
F(2, 65)	22645.67	P-value(F)	3.79e-93
Log-likelihood	-398.1071	Akaike criterion	802.2142
Schwarz criterion	808.8728	Hannan-Quinn	804.8525
rho	0.516487	Durbin-Watson	0.810489

Statistics based on the original data:

Mean dependent var	149.4950	S.D. dependent var	65.69289
Sum squared resid	390.4095	S.E. of regression	2.450775

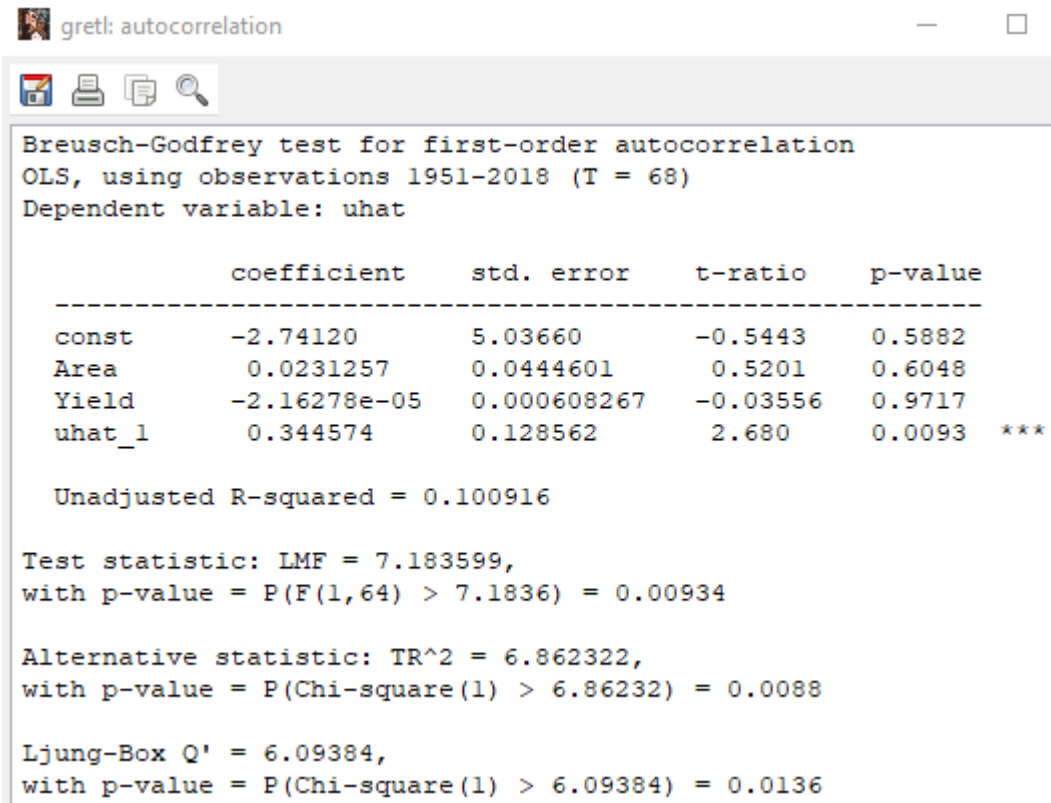
We perform Weighted Least Squares from other linear models and use the independent variable Yield as the weight. However, gretl does not allow to test for heteroskedasticity in this model so we leave it as it is.

3) Autocorrelation- autocorrelation is the correlation between values of the process at different points in time, as a function of the two times or of the time difference.

Testing Autocorrelation

H₀: No Autocorrelation

H_A: Autocorrelation is present



```
gretl: autocorrelation

Breusch-Godfrey test for first-order autocorrelation
OLS, using observations 1951-2018 (T = 68)
Dependent variable: uhat

      coefficient    std. error    t-ratio    p-value
-----
const    -2.74120      5.03660     -0.5443    0.5882
Area      0.0231257      0.0444601     0.5201    0.6048
Yield    -2.16278e-05     0.000608267   -0.03556   0.9717
uhat_1     0.344574      0.128562      2.680     0.0093 ***

Unadjusted R-squared = 0.100916

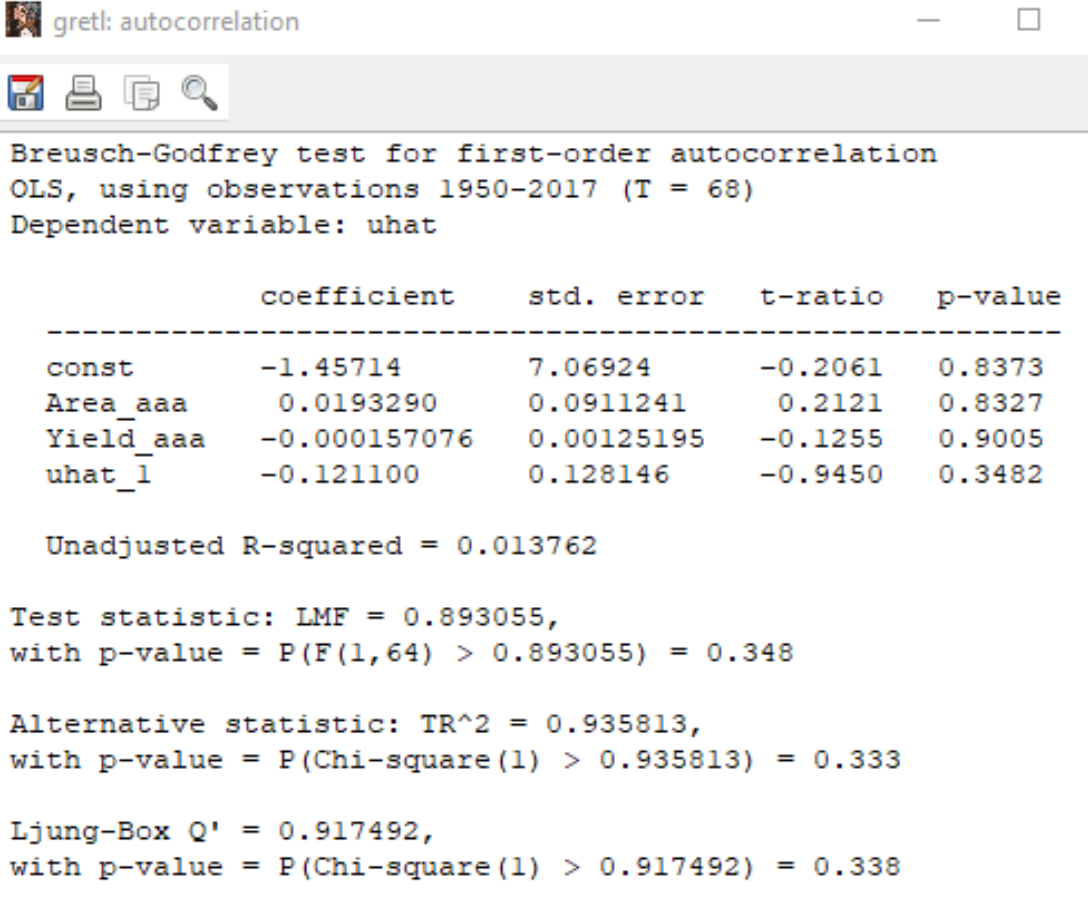
Test statistic: LMF = 7.183599,
with p-value = P(F(1,64) > 7.1836) = 0.00934

Alternative statistic: TR^2 = 6.862322,
with p-value = P(Chi-square(1) > 6.86232) = 0.0088

Ljung-Box Q' = 6.09384,
with p-value = P(Chi-square(1) > 6.09384) = 0.0136
```

Since $p\text{-value} = P(F(1,64) > 7.1836) = 0.00934$ is less than 0.05 (general level of significance) so reject null hypothesis, so autocorrelation is present in our data.

Remedy for Autocorrelation -



The image shows a screenshot of the 'gretl: autocorrelation' window. It displays the results of a Breusch-Godfrey test for first-order autocorrelation. The test is based on OLS using observations from 1950 to 2017 (T = 68). The dependent variable is 'uhat'. A table of coefficients, standard errors, t-ratios, and p-values is shown. Below the table, the unadjusted R-squared is 0.013762. Test statistics for LMF, TR^2, and Ljung-Box Q' are also provided, all with p-values greater than 0.3, indicating no significant autocorrelation.

```
gretl: autocorrelation

Breusch-Godfrey test for first-order autocorrelation
OLS, using observations 1950-2017 (T = 68)
Dependent variable: uhat

      coefficient    std. error    t-ratio    p-value
-----
const      -1.45714      7.06924     -0.2061    0.8373
Area_aaa     0.0193290     0.0911241     0.2121    0.8327
Yield_aaa   -0.000157076    0.00125195   -0.1255    0.9005
uhat_1      -0.121100      0.128146    -0.9450    0.3482

Unadjusted R-squared = 0.013762

Test statistic: LMF = 0.893055,
with p-value = P(F(1,64) > 0.893055) = 0.348

Alternative statistic: TR^2 = 0.935813,
with p-value = P(Chi-square(1) > 0.935813) = 0.333

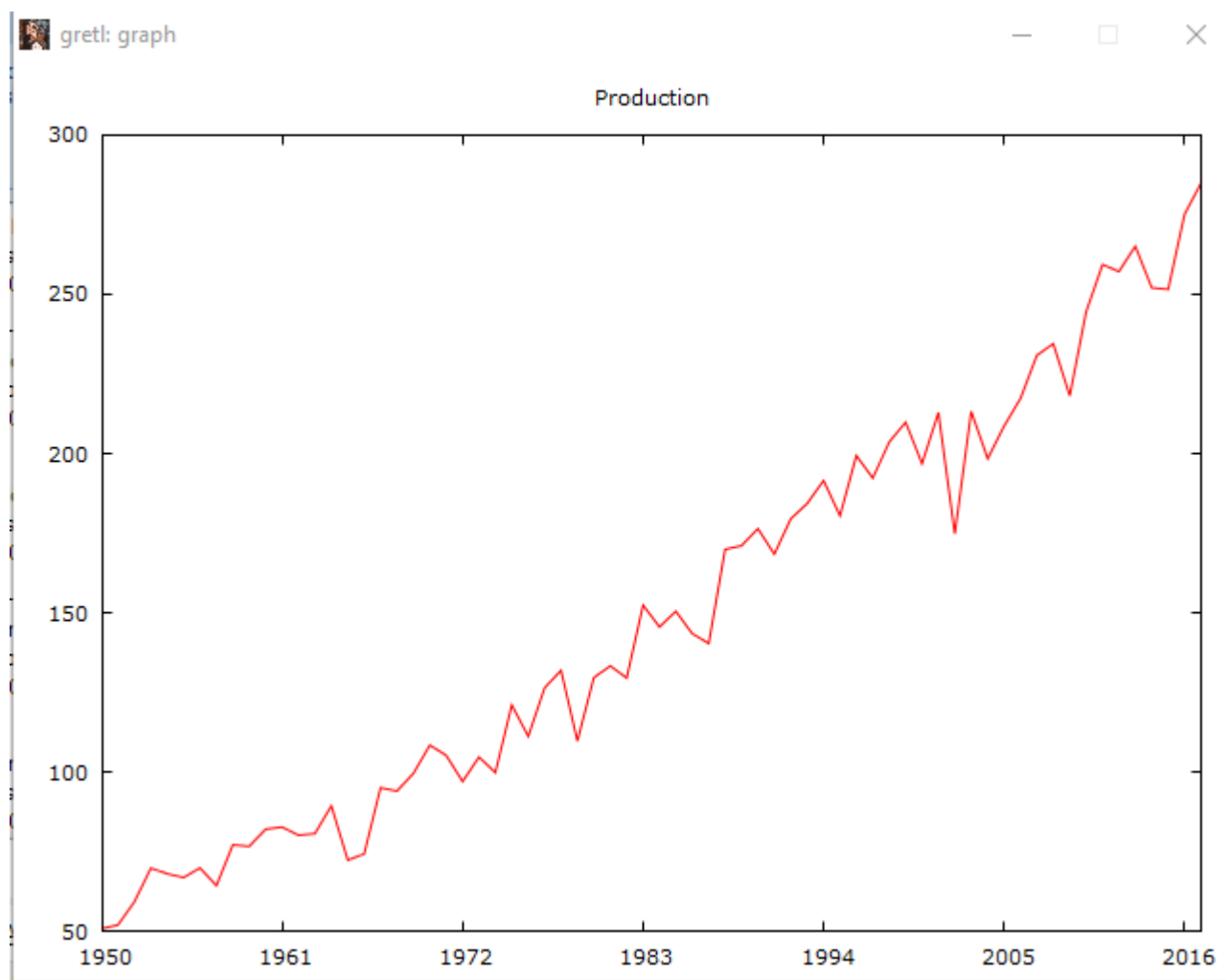
Ljung-Box Q' = 0.917492,
with p-value = P(Chi-square(1) > 0.917492) = 0.338
```

After using Prais-Winston transformation under GLS, I was able to remove autocorrelation from my model.

Empirical Results

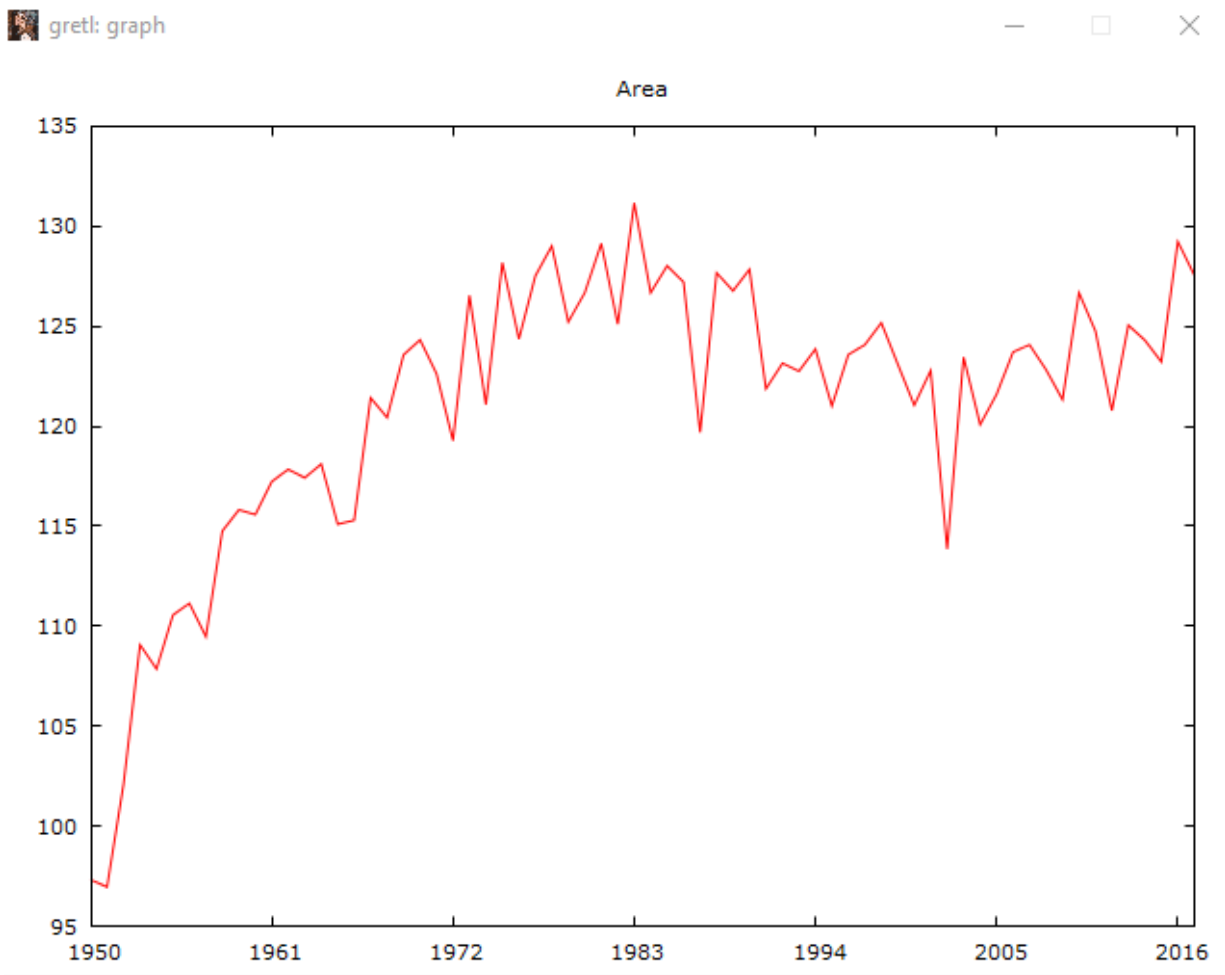
From the following graph (Graph 1) (INDIA 1950-2016) we can conclude that Production of Foodgrains in India has seen almost a 5 fold increase in the last six decades, with the sharpest drop around 2008 near the time of the global financial crisis of 2008.

Graph 1



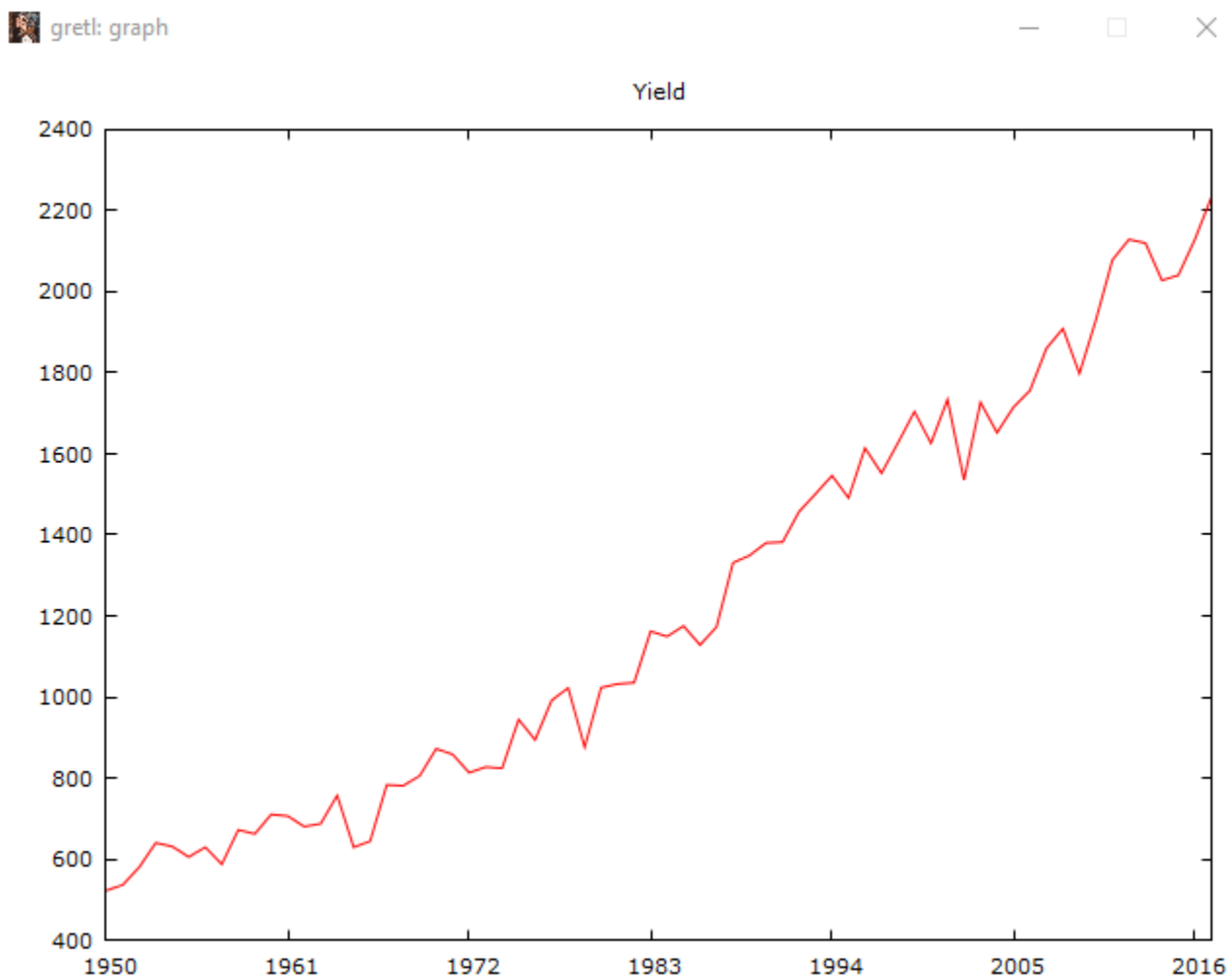
Graph 2 shows that there has been a drastic increase in the area of land used for production of foodgrains over the last six decades by approximately 30 million hectares with sharp falls in 1985 and 2005 which might be due to shift from agricultural sector to industrial and corporate sectors as a result of globalization and industrialization.

Graph 2

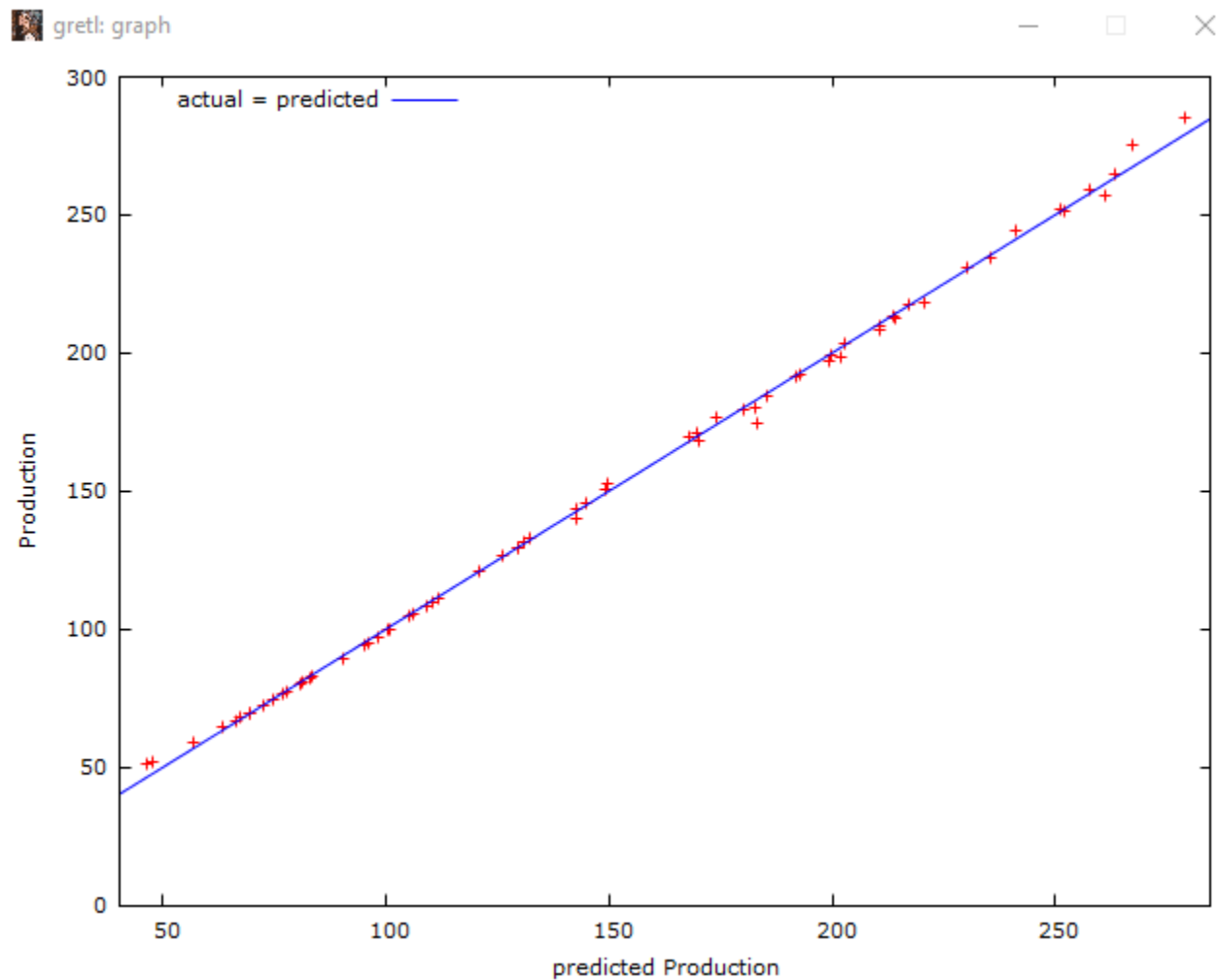


Graph 3 shows that per kg Yield from per hectare of land has increased by more than 4 times. Much of this increase can be credited to the Green Revolution started by Mr. M.S. Swaminathan in mid 1960's. The sharp fall in production in 2003-04 can be attributed to the excessive rains, floods and droughts in that year.

Graph 3



Actual VS Fitted (Production of Foodgrains)



The Model predictions are quite Accurate as from the above graph we can see that there are very less scattered values this shows there's a strong correlation between the model's predictions and its actual values regarding the Production of Foodgrains.

Model 3: OLS, using observations 1950-2017 (T = 68)

Dependent variable: Production

	coefficient	std. error	t-ratio	p-value	
const	-89.5747	5.16091	-17.36	4.29e-026	***
Area	0.735376	0.0456423	16.11	2.20e-024	***
Yield	0.123015	0.000636487	193.3	1.94e-091	***
Mean dependent var	149.4950	S.D. dependent var	65.69289		
Sum squared resid	331.0111	S.E. of regression	2.256652		
R-squared	0.998855	Adjusted R-squared	0.998820		
F(2, 65)	28356.65	P-value (F)	2.56e-96		
Log-likelihood	-150.2977	Akaike criterion	306.5954		
Schwarz criterion	313.2540	Hannan-Quinn	309.2338		
rho	0.327389	Durbin-Watson	1.244508		

This OLS regression shows following results:

- 1) Coefficient of Constant tells that if the value of ProdR, ProdW and AreaF will be 0 then ProdF would be -89.5747 which does not make any economic sense.
- 2) One Unit increase in the Area keeping Yield constant would lead to 0.735376 units increase in the level of Production.
- 3) One unit increase in the Yield keeping Area constant would lead to 0.123015 units of increase in the level of Production.
- 4) All t ratio's (All $|t| > 2$) are significant suggesting that coefficients are statistically significant (so all independent variables have significant impact on dependent variable)
- 5) As the value of R-square is 0.99885 that means 99.85 % of the variation in ProdF is explained by ProdW, ProdR, AreaF.
- 6) Only 0.15 % of the variation is left unexplained

8) Coefficients are also economically significant as their sign is as expected from the economic theory. Positive relation between Area, Yield and Production matched with the positive signs of the coefficient.

Conclusion

In the present work the relationship between Dependent Variable- Production of Foodgrains and Explanatory Variables- Area for Production of Foodgrains and Yield per hectare of land (kg/hectare units) has been examined. We observed that the level of Production of Foodgrains has increased over the last 60 years in the India and it's movement is closely in resemblance with increase in Yield of land. The study shows that both the explanatory variables had significant effect on the dependent variable. Areas under production of foodgrains, Yield per hectare of land are both positively related to the Production of foodgrains which is in accordance with the economic theory.

Violation of OLS Assumptions was also tested in this study and was found that the initial model had heteroskedasticity and autocorrelation violations. (Autocorrelation corrected by Prais-Winstone transformation under GLS)

Overall, the empirical results seemed to support that Production of Foodgrains is being significantly affected by the dependent variables. But it is acknowledged that there can be several other underlying factors like water availability, seasonal changes and monetary incentives for farmers adding to the disparity which were beyond the scope of this study.

