



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Robust regression model for prediction of Soybean crop price based on various factors

¹K.Karthikeyan, ²Akshay Harlalka

¹Associate professor, SAS, Mathematics Division, VIT University, Vellore-14, Tamil Nadu, India,

²SMBS, VIT University, Vellore-14, Tamil Nadu, India.

Abstract: Prediction of future food prices involves taking a lot of critical factors like temperature, precipitation and crop yield in to consideration. Most regression models consider only climatic and scientific factors in to consideration which give only one side of the picture. This article gives a more balanced overview as it considers a host of other factors which indirectly play a significant role in crop prices especially the economic factors. Multi-linear regression model is used to predict the price of Soybean in USA during the 11 year period from 1995 to 2005 and to compare the factors affecting food price. This model explain more than 90% of the variation in the crop price based on just four major selected factors and shown that there is a very strong relationship between observed values and model predicted values with a multiple correlation coefficient of 0.949 for USA. Also we use the F-test to test the significance of the regression relationship between crop prices and selected factors.

Keywords: Crop prices, Regression model, Correlation, Partial regression

I. INTRODUCTION

The recent crisis in food prices, which has affected thousands of families throughout the developing world, has once again underscored the urgent need for governments to strengthen their safety net systems to ensure that the rise in the price of basic commodities does not trigger an increase in poverty rates. Jordan Schwartz, World Bank lead economist for sustainable development in Latin America and the Caribbean, mentioned several factors that are driving the price increase: speculation in commodity markets, the booming demand from Asia for feed grains and land use switching out from food crops to biofuels, among others. There is growing consensus that food prices have increased due to fundamental shifts in global supply and demand. A variety of forces contribute to rising food prices: high energy prices, increased income, climate change and the increased production of biofuel. Income and per capita consumption in developing countries has increased; consequently, demand has also risen. Changes in food supply and demand have been accompanied by predictable effects in terms of pricing and have been further affected by the rise in the cost of non-renewable resources[1]. There are many systematic studies being done in various countries on the prediction model of different crops. But a majority of studies have taken only the influence of climate change on crop prices into consideration. Nicholls[2] estimated the contribution of climate trends in Australia to the substantial increase in Australian wheat yields since 1952. Non-climatic influences such as new cultivars and changes in crop management practices were removed by detrending the wheat yield and climate variables and using the residuals to calculate quantitative relationships between variations in climate and yield. Lobell[3] used a combination of mechanistic and statistical models to show that much of this increase in wheat yields in Mexico can be attributed to climatic trends in Northwest states, in particular cooling of growing season nighttime temperatures. Despite the complexity of global food supply, Field[4] showed that simple measures of growing season temperatures and precipitation spatial averages based on the locations of each crop explain 30% or more of year-to-year variations in global average yields for the world's six most widely grown crops. For wheat, maize and barley, there is clearly negative response of global yields to increased temperatures. Burke[5] used a perfect model approach to examine the ability of statistical models to predict yield responses to changes in mean temperature and precipitation, as simulated by a process-based crop model. Kaufmann[6] estimated a model that accounts for both climatic and social determinants of corn yield in the United States. Climate variables are specified for periods that correspond to phenological stages of development. Social determinants include market conditions, technical factors, scale of production, and the policy environment. Bonfils[7] concluded that climate change in California is very likely to put downward pressure on yields of almonds, walnuts, avocados, and table grapes by 2050. Without CO₂ fertilization or adaptation measures, projected losses range from 0 to greater than 40% depending on the crop and the trajectory of climate change. Climate change uncertainty generally had a larger impact on projections than crop model uncertainty, although the latter was substantial for several crops. Lobell[8] seeks to improve quantitative understanding of price spikes in general and the potential effects of climate change on these spikes in particular. Naylor[9] provided an insight into the causes and consequences of the volatile events like the 2008 food price run up. Naylor mentioned that price variability, particularly spikes, has enormous impacts on the rural poor who spend a majority of their income on food and have minimal

savings. Impacts at the local level have not been well measured, yet are the key to improving food security globally. To see how the rural poor were impacted on a local scale, Naylor and Falcon looked at Ghana, Uganda, Malawi, Guatemala, and India. Price changes at the local level during the 2008 price spike were frequently half that of international prices, primarily as a consequence of domestic food and trade policies. Additionally, domestic self-sufficiency policies tended to have long-term negative impacts on the international market when governments lacked the resources to defend a targeted price or were large actors with significant shares of global production or consumption.

II. PRELIMINERIES

Linear regression is used to model the value of a dependent scale variable based on its linear relationship to one or more predictors. The linear regression model assumes that there is a linear or straight line relationship between the dependent variable and each predictor. This relationship is described in the following formula.

$$y_i = b_0 + b_1X_{i1} + b_2X_{i2} + b_3X_{i3} + \dots + b_pX_{ip} + e_i$$

where

- y_i - the value of the i^{th} case of the dependent scale variable
- p - the number of predictors
- b_j - the value of the j^{th} coefficient, $j = 0, \dots, p$
- x_{ij} - the value of the i^{th} case of the j^{th} predictor
- e_i - the error in the observed value for the i^{th} case

The model is linear because increasing the value of the j^{th} predictor by 1 unit increases the value of the dependent by b_j units. Note that b_0 is the intercept the model-predicted value of the dependent variable when the value of every predictor is equal to 0.

For the purpose of testing hypotheses about the values of model parameters, the linear regression model also assumes the following:

- The error term has a normal distribution with a mean of 0.
- The variance of the error term is constant across cases and independent of the variables in the model. An error term with non-constant variance is said to be heteroscedastic.
- The value of the error term for a given case is independent of the values of the variables in the model and of the values of the error term for other cases.

Before, using linear regression, the correlation between each of the independent variables and the dependent variable must be studied to determine whether a linear model is suitable for those variables. This study uses the multiple regressions modeling to predict the crop prices of Soybean crop in USA. The relationship between the variables and crop prices is studied systematically using the above procedure and a final regression model with four most important variables is created for each case.

III. DATA SOURCES

Eight major factors were identified and the corresponding information was collected from reliable sources. The following data on the crop prices, import and export quantity, total population, crop yield, and food supply were obtained from the Food and Agriculture Organization of the United Nations[11]. The average annual temperature and precipitation data of United States of America was obtained from the National Climatic Data Center of NOAA (National Oceanic and Atmospheric Administration)[12]. The crude oil prices were obtained from the Energy Information Administration of US Department of Energy[13]. The data table explains the variation of crop prices with regard to eight different variables. USD/bbl refers to US Dollar per barrel. Hg/ha refers to hectograms per hectare. The representation of the symbols is as given below:

CP refers to Crop Price (USD/tonne); **COP** refers to Crude Oil Price (USD/bbl)

Total P refers to Total Population (in thousands); **Temp** refers to Temperature (in Fahrenheit)

Ppt refers to Precipitation (in inches); **ExQ** refers to Export Quantity (tonnes)

ImQ refers to Import Quantity (tonnes); **FS** refers to Food Supply (tonnes); **CY** refers to Crop yield (Hg/ha)

DATA TABLE

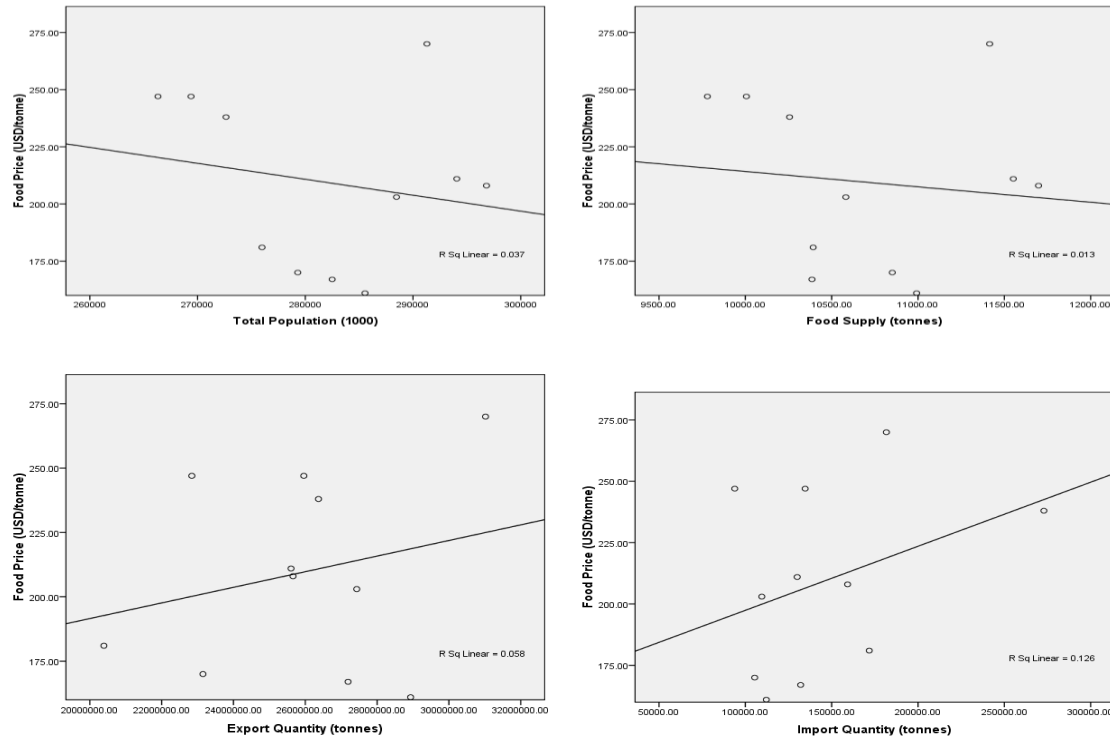
CP	COP	Total P	Temp	Ppt	ExQ	ImQ	FS	CY	Year
247	14.62	266324	52.71	31.69	2.28E+07	134644	9780	23758.93	1995
247	18.46	269394	51.89	32.59	2.60E+07	93847	10006	25269.93	1996
238	17.23	272643	52.26	31.29	2.64E+07	272900	10256	26165.48	1997
181	10.78	275986	54.32	32.97	2.04E+07	171757	10393	26168.84	1998
170	15.56	279300	53.93	27.84	2.32E+07	105397	10851	24634.35	1999
167	26.72	282496	53.27	27.73	2.72E+07	132025	10386	25613.18	2000

161	21.84	285545	53.68	28.58	2.89E+07	112128	10992	26638.6	2001
203	22.51	288467	53.21	28.66	2.74E+07	109517	10582	25566.31	2002
270	27.56	291291	53.29	29.95	3.10E+07	181673	11415	22768.74	2003
211	36.77	294063	53.12	32.88	2.56E+07	130032	11552	28404.15	2004
208	50.28	296820	53.64	29.84	2.57E+07	159197	11698	28959.96	2005

IV. RESULTS AND DISCUSSION

A. Strength of relationship between variables

The variables involved in the following scatter plots were excluded from the prediction model since their correlation with the crop price was insignificant.



From the analysis, it was concluded that temperature is the best predictor of food prices in USA, followed by precipitation and crop yield. All other factors have insignificant influence on the crop prices and hence they shall be excluded from the prediction model.

B. Regression Model for Prediction

The prediction model is

$$CP(t) = b_0 + b_1Temp(t) + b_2Ppt(t) + b_3CY(t) + b_4COP(t) \quad \text{---(I)}$$

where CP- Crop Price (USD/tonne), COP- Crude Oil Price (USD/bbl), Temp- Temperature (in Fahrenheit), Ppt- Precipitation (in inches), CY- Crop yield (Hg/ha)

C. Robustness of Prediction model

The following is the model summary and ANOVA table (obtained by the SPSS software):

ANOVA(TABLE 3)

Model	Sum of Squares	Df	Mean Square	F	Sig.
1 Regression	12480.231	4	3120.058	13.543	.004 ^a
Residual	1382.315	6	230.386		
Total	13862.545	10			

a. Predictors: (Constant), Crude Oil Price (USD/bbl), Temperature (farhenheit), Precipitation (inches), Crop Yield (Hg/ha)

b. Dependent Variable: Food Price (USD/tonne)

F_model for Regression

Suppose that the regression assumption hold and that the linear regression model has $k+1$ parameters and consider testing $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$; H_a : Atleast one of $\beta_1, \beta_2, \beta_3, \beta_4$ does not equal to zero.

$$F(model) = \frac{(Explained\ variation)/k}{(Unexplained\ variation)/(n-(k+1))} = 13.543$$

where n is the number of observations and k is the number of independent variables. The significance value of the F statistic is less than 0.05, which means that the variation explained by the model is not due to chance. This analysis of variance table gives us a first-hand proof that the model considered is suitable for multi- linear regression.

Model summary (TABLE 4)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.949 ^a	.900	.834	15.17846	.900	13.543	4	6	.004

a. Predictors: (Constant), Crude Oil Price (USD/bbl), Temperature (farhenheit), Precipitation (inches), Crop Yield (Hg/ha)

b. Dependent Variable: Food Price (USD/tonne)

Multiple coefficient determination (R^2) and Adjusted multiple coefficient of determination (Adjusted R^2)

$$R^2 = \text{Explained variation} / \text{Total variation} = 12480.231 / 13862.545 = 0.9002843$$

$$\text{Adjusted } R^2 = (R^2 - (k/(n-1))) * ((n-1)/(n-(k+1))) = 0.834$$

R , the multiple correlation coefficient tells us how much accurate a model is in predicting the crop prices. R has a value of 0.949 which indicates a very strong relationship between observed and the model predicted values of the dependent variable. R Square, the coefficient of determination indicates that this model explain about 90% variation in the crop prices. This shows that the model accounts for a majority of the variation in the crop prices. As a further measure of strength of the model fit, the standard error of the estimate of the model (15.17846) is considerably lower than the standard deviation of the crop price. (37.23244). This also measures the accuracy of the model in predicting the crop prices.

D. Regression and Correlation Coefficients

Regression Coefficients (TABLE 6)

Model	Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
Constant (b_0)	$b_0 = 1340.257$	418.054		3.206	.018		
Temperature (farhenheit)	$b_1 = -20.069$	7.477	-.384	-2.684	.036	.813	1.230
Precipitation (inches)	$b_2 = 10.809$	2.757	.584	3.920	.008	.748	1.338
Crop Yield (Hg/ha)	$b_3 = 0.017$	.004	-.835	-4.634	.004	.512	1.954
Crude Oil Price (USD/bbl)	$b_4 = 2.169$	.570	.662	3.803	.009	.549	1.822

a. Dependent Variable: Food Price (USD/tonne)

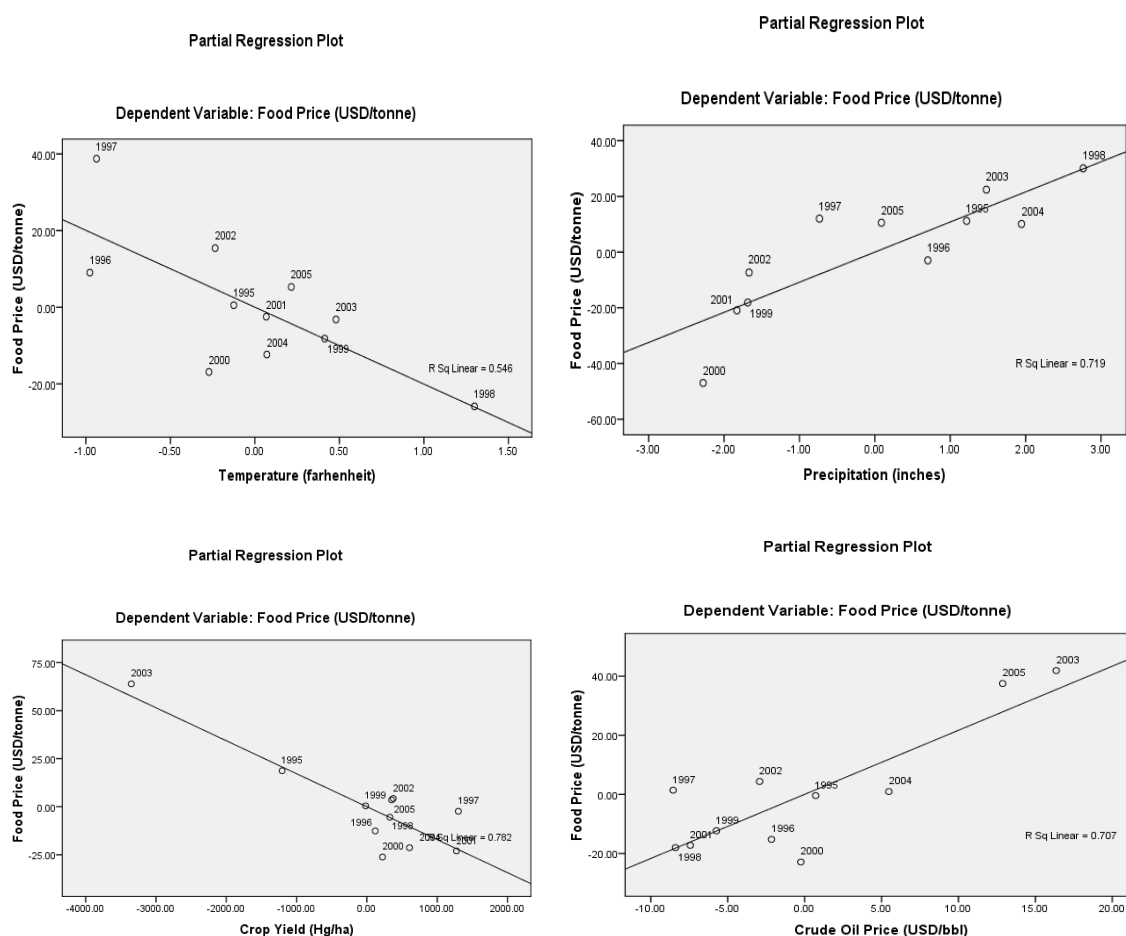
Refer Equation no. I

The Sig. and Standardized Coefficients column show that all the factors are quite significant predictors of the food/crop price. This shows that the crop yield (Hg/ha) contributes the most to the model as it has the highest standardized coefficient. The tolerance values are significant indicating that there is low multi-collinearity and the more than 50% of the variance in a predictor cannot be explained by other variables. This shows that the regression model is very much accurate with low standard error of the regression coefficients. The Variance Inflation Factor of all the variables is less than 2 which confirms the previous statement.

E. Partial regression plots

A partial regression plot is a scatterplot of the partial correlation of each independent variable with the dependent variable after removing the linear effects of the other independent variables in the model. Each plot is considered to see if it shows a linear or nonlinear pattern. If the specific independent variable shows a linear relationship to the dependent variable, it meets the linearity assumption of multiple regression[16]. These partial

regression plots show a linear relationship between the dependent and specific independent variables. Hence, this satisfies the linearity assumption of the multiple regression.



Summary of Findings

Soybean Crop Prices	Findings
USA	1. Climatic factors play a major role in the prediction of prices 2. Role of economic factors like Import Quantity, Export Quantity not significant Reason: Soybean is a widely grown crop in USA and one of the largest exporters of the crop. Fluctuations in climate can cause a huge shift in the supply and demand causing price variations.

V. CONCLUSION

Temperature is the most important predictor of soybean crop prices for USA. Climatic factors accounted for much of the variation in the food prices in USA. The most of the food price model do not take economic factors into account thus compromising the accuracy and robustness of the model. The prediction model account for more than 90% variation in crop prices and a strong relationship has been established between the observed and model predicted values with a multiple correlation coefficient close to 0.95.

ACKNOWLEDGEMENTS

We would like to thank the Food and Agriculture Organization of the United Nations, National Climatic Data Center of NOAA (National Oceanic and Atmospheric Administration) and Energy Information Administration of US Department of Energy for providing the relevant information for research and analysis.

REFERENCES

- [1]. World Bank News, Feature Story, September 13, 2012 What are the facts about rising food prices and their effect on the region?
- [2]. Nicholls N 1997 Increased Australian wheat yield due to recent climate trends, Nature, Vol.No. 387, pp. 484–485.
- [3]. Lobell D B, Ortiz-Monasterio J , Asner G P, Matson P A, Naylor R L and Falcon W P 2005 Analysis of wheat yield and climatic trends in Mexico Field Crops Res. 94 250–6
- [4]. David B Lobell and Christopher B Field, Global scale climate–crop yield relationships and the impacts of recent warming, Environ. Res. Lett. 2 (2007) 014002 (7pp)

- [5]. David B.Lobella, MarshallB.Burke, On the use of statistical models to predict crop yield responses to climate change, Article in press, 2010
- [6]. Robert K. Kaufmann, A Biophysical Model of Corn Yield: Integrating Climatic and Social Determinants, American Journal of Agricultural Economics, Volume 79, Issue 1, 178-190.
- [7]. Lobell D B, Field C B, Cahill K N and Bonfils C 2006 Impacts of future climate change on California perennial crop yields: model projections with climate and crop uncertainties Agric Meteorol. 141 208–18
- [8]. David Lobell, Food Price Spikes in a Warming World, FSI Stanford, FSE Project 2010-Present. (http://foodsecurity.stanford.edu/research/food_price_spikes_in_a_warming_world/)
- [9]. Rosamond L. Naylor, Food price volatility and the world's poor, January 3, 2011 - FSI Stanford, FSE News, (http://foodsecurity.stanford.edu/news/food_price_volatility_and_the_worlds_poor_20110103/)
- [10]. Nomura Global Economics, Global Economics and Strategy, September 2010
- [11]. FAO (Food and Agriculture Organization of the United Nations) FAO Statistical Databases available from: <http://faostat.fao.org>
- [12]. National Climatic Data Center of NOAA (National Oceanic and Atmospheric Administration) Database available from: <http://www.ncdc.noaa.gov/cag/>
- [13]. US Department of Energy, Energy Information Administration, Petroleum Marketing Annual 2009, DOE/EIA-0487(2009)
- [14]. Multi collinearity www.chsbs.cmich.edu/fattah/courses/empirical/multicollinearity.html
- [15]. University of Texas, Illustration of Regression analysis, http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/045_Linearity_of_Independent_Variables_Partial_Plots.html