Association between Food Production and Economics Statistics

1. Introduction

1.1 Statement of the research

The research studies the association between food production and the economics statistics. It will use the food production index as the response variable and the explanatory variables include national GDP, food imports and exports %GDP, and consumer price inflation. The research analyzes both the geographic and social factors which impact the food production of different countries.

1.2 Motivation of the research

The food production concerns people across all the countries. By identifying the factors in contribution of the food production, the research provides a fundamental view of food safety and capacity. The geographic factors include forest land, farmable land and water safety. The social factors include the birth rate, life expectancy and industry structure.

1.3 Implications of the research

The research will help us find the ways to improve food production in the world. Furthermore, we will be able to analyze how industrial structure impact the food production and how living styles shape the food producing activities. The research will make us aware of the food related issues and get more involved in enhancing the food safety and stability around us.

2. Methods

2.1 Sample

The World Bank codebook is used in the research. The data was collected as of March 2016 and the data period is 2012 – 2013. It includes all the countries' GDP, population, geography, and industry information. The sample are selected based on the country records and the food production related factors. There are 248 observations and 163 variables in the data codebook. In the research, 22 variables are selected for all the 248 observations to represent each countries' conditions.

2.2 Measures

1) Description of variables

There are 22 explanatory variables in the sample, including geographic parameters (agricultural land, population growth etc) and social parameters (GDP and its growth, imports and exports, industrial structures etc). The response variable is the food production index. As both explanatory variables and response variable are quantitative base, we will conduct a quantitative – quantitative analysis by using model training methods.

2) Manage variables

In the research, we convert all the quantitative variables into float base and replace null data with 0 to make the data analysis more accurate. Year 2012 data are selected to analyze each variable. *(table1)*

2.3 Analytics

The following steps are used for the analysis:

i) Data clean and preparation

ii) Analyze the mean/standard deviation

iii) Visualize the variables by scatter plots

iv) Conduct factor analysis by Lasso Regression

v) Check the data results by cross validation

We clean the data sample by ruling out the null records and converting each variable to numerical type. We set 22 explanatory variables as predictors and the response variable as the target. 70% of the dataset is selected as training data and 30% as testing data.

Firstly, we use the pandas package to calculate the mean/ standard deviation/ quantile for each variables. It gives as a preview of the data set distribution. Then, we use scatter plots to visualize the data distribution and cluster the variables. We employ the linear regression model to describe all the factors contributing to the food production index. By using the Lasso Regression analysis, we can identify the key factors and train a model to describe the correlation between food production and its explanatory variables. Finally, we use the Cross Validation to exam the data selection and analysis results. The sklearn tool kits are applied for the analysis. The data visualization, factors clustering and model prediction.

3. Results

3.1 Descriptive Analysis

In our hypothesis, we assume that the food production index should be correlated with the GDP annual growth rate. To take a further look of the correlation, we make the bar plot to show the food production index distribution as of the GDP annual growth. We divide the GDP annual growth of each country into four categories sorting in ascending order (1=25%, 2=50%, 3=75%, 4=100%). In the bar plot as below, we see the food production index distribution has obvious variations in the quantiles of GDP annual growth. More specifically, the food production index increases with the GDP annual growth. In the 25% quantile of GDP growth the index is 100, while in the 75% quantile of GDP growth the index is around 125. *(figure1)*

In the research, we select 22 variables to study the factors impacting the food production index. We use pandas data describe function to analyze the mean, max, min, and standard deviation of each variable. The following table shows the representative analysis of these variables: *(table2)*

The mean of exports of goods and services is 43.97% of GDP, with max 225.56% and min 5.52%. The standard deviation of exports is 28.89%, which indicates the large variance. The mean of imports of goods and services is 49.55% of GDP, with max 224.43% and min 12.94%. The stand deviation of imports is 26.53%. The exports and imports of goods and services show the same distribution pattern.

The average GDP at market prices is $2,206 billion, with max $74,200 billion and min $0.039 billion. The SD is $8,205 billion, which indicates that the GDP at market prices vary largely among different countries.

3.2 Bivariate Analysis

To study the bivariate relationship between food production and each factor, we use scatter plots to analyze the correlation between two variables. As most of the variables are quantitative including the response variable, we then use the Pearson Correlation to analyze distribution of food production based on each factor.

From the scatterplot for exports goods and service, we see the food production distribution is concentrated in 50-100 exports goods and services. Furthermore, the food production distribution leans to downward to show a negative relationship between the food production and exports of goods and services. Similarly, the imports of goods and services show the same pattern of distribution. Among the countries, the exports and imports increase, while the food production index decreases.

In the Scatterplot 3, the GDP annual growth increases while the food production index increases. The food distribution is more concentrated, but still shows a positive correlation with GDP annual growth. In the Scatterplot 4, the urban population shows a positive impact on the food production index as well. The mode of distribution is between 0-6% population growth rate per year. *(figure2-5)*

To identify the coefficient and its significance-level, we use the Pearson Correlation to analyze the quantitative to quantitative relationship. The correlation and p-value are summarized as below: *(table3)*

According to the table, the coefficient of exports goods & services is -0.419 with p-value of 0.0014 (< .05). Hence, we reject the H0 hypothesis and say that exports has significant impact on the food production. Comparably, the coefficient of imports is -0.235 with p-value of 0.073 (>.05). The null hypothesis is accepted, and imports doesn't correlate to the food production significantly.

We conclude that exports goods and services, GDP annual growth, and urban population growth all have significant correlation with the food production, while the imports goods and services impact is not significant. The coefficients for the first three factors are -0.419, 0.587, and 0.697 accordingly.

3.3 Multivariable analysis

Lasso Regression

In the bivariate analysis session, we already identify three correlation between food production and 3 out of 4 explanatory variables. In this session, we are going to use Lasso regression and cluster analysis to validate all the 22 variables. The linear regression is used to build the food production factoring model. We can identify 5 out of 22 explanatory variables as main contribution to the food production index. Out test data takes up 30% of the total dataset. The MSE of training dataset is 193 and MSE of test dataset is 199, which shows consistent of mean square error in both the datasets. The R-square of training dataset is 0.616 and the R-square of test dataset is 0.469. The linear model explains 61.6% of variance in the training dataset and 46.9% of variance in the test dataset.

The following table shows the betas and names of each variable: *(table4)*

We then run the regression coefficient and mean squared error for 10 folds of the model simulation. The following figures show that variance of the variables' errors are reduced after the k-fold Lasso regression. The 22 explanatory variables are well described by the linear model. *(figure6-7)*

Cluster Analytics

By clustering the factors, we expect to see some of the explanatory variables to be categorized into the same cluster. We use the elbow method to calculate the average distance of the variables as the number of clusters increase. The average distance is reduced effectively after the clusters of variables are raised to 3.

We use the canonical discriminant method to draw the variance distribution based on the three clusters hypothesis. From the scatterplot, we can see the variances of the variables are allocated in three groups, which support our assumption. In the cluster analysis, we breakdown the factors of food production index and then group them into three categories. *(figure8-9)*

Next step, we want to exam the significance of the three clusters by using the ANOVA analysis. The variances of the explanatory variables are standardized to mean of 0 and standard deviation of 1. The Cluster 0 has positive correlation for one variable (X140_2012) and Cluster 2 has positive correlation for four variables (X140_2012, X162_2012, X284_2012, and X35_2012). In contrast, Cluster 1 has negative correlation for the last four variables. *(table5)*

Furthermore, we want to exam how well the three clusters of variables can explain the response variable of food production index (X129_2012). The probability of F-statics is <.05 indicating that correlation between the three clusters and X129_2012 is significant.

The means of X129 in the three clusters vary significantly and the standard deviation of X129 also change by each cluster: *(table6-7)*

The multiple comparison between the three clusters shows that the mean difference is significant. Therefore, we can conclude that the response variable can be explained by the three clusters of factors. *(table8)*

4. Conclusions/Limitations

4.1 Overview

In the research, we analyze the factors that impact the food production index. We finally identify 5 out of 22 variables to explain the response variable of food production. In the bivariate analysis, we find that the exports/imports of goods and services are negatively correlated with food production, while the GDP growth and urban population growth are both positively correlated with food production. The Lasso regression and Cluster analytics are used in the multivariate analysis. We group the factors into three categories and over 65% of the training dataset can be explained by the selected factors.

4.2 Implications

We find that five factors impacting the food production index: the exports of goods and services (beta= -0.1659), the GDP growth (beta= 1.3652), the inflation consumer prices (beta=3.2878), the urban population growth (beta=5.8646), and the agriculture value added (beta= 6.8306). To make food production more effective, we can improve the five factors and create a dynamic metrics to measure the food industry growth. We can also focus more on the agriculture value added in driving the food production improvement.

4.3 Limitations/Future Direction

Only Year 2012 is selected for the data sample. The parameters may vary from year to year. Hence, a wider time of period can be introduced to have a more comprehensive view of the global impacts of food production. Also, the other factors may have a bigger impact on some of the countries, but the impact level is not significant for all the countries. In future study, a longer period and more detailed factoring can be included.

Appendix

Table 1:

Variable Label	Variable Name
X121_2012	EXPORTS OF GOODS AND SERVICES (% OF GDP)
X134_2012	FOREST AREA (% OF LAND AREA)
X139_2012	GDP AT MARKET PRICES (CURRENT US$)
X140_2012	GDP GROWTH (ANNUAL %)
X142_2012	GDP PER CAPITA (CURRENT US$)
X143_2012	GDP PER CAPITA GROWTH (ANNUAL %)
X154_2012	IMPORTS OF GOODS AND SERVICES (% OF GDP)
X156_2012	IMPROVED WATER SOURCE (% OF POPULATION WITH ACCESS)
X161_2012	INDUSTRY, VALUE ADDED (% OF GDP)
X162_2012	INFLATION, CONSUMER PRICES (ANNUAL %)
X171_2012	LIFE EXPECTANCY AT BIRTH, FEMALE (YEARS)
X174_2012	LIFETIME RISK OF MATERNAL DEATH (%)
X179_2012	MANUFACTURING, VALUE ADDED (% OF GDP)
X195_2012	NET MIGRATION
X220_2012	POPULATION GROWTH (ANNUAL %)
X258_2012	RURAL POPULATION (% OF TOTAL POPULATION)
X283_2012	URBAN POPULATION (% OF TOTAL)
X284_2012	URBAN POPULATION GROWTH (ANNUAL %)
X31_2012	AGRICULTURAL LAND (% OF LAND AREA)
X35_2012	AGRICULTURE, VALUE ADDED (% OF GDP)
X36_2012	AGRICULTURE, VALUE ADDED (ANNUAL % GROWTH)
X45_2012	ARABLE LAND (% OF LAND AREA)
X9_2012	ADJUSTED NET NATIONAL INCOME (CURRENT US$)

Table 2:

Factors Label	X121_2012	X154_2012	X140_2012	X162_2012	X139_2012
count	209.00	209.00	222.00	207.00	225
mean	43.97	49.55	3.66	5.22	$2,206,687,000,000
std	28.89	26.53	8.41	5.15	$8,205,548,000,000
min	5.52	12.94	-46.08	-0.94	$39,875,750
25%	26.41	31.46	1.11	2.51	$9,359,185,000

50%	37.71	43.97	3.28	3.73	$45,300,670,000
75%	53.81	60.63	5.27	6.24	$397,000,000,000
max	225.56	224.43	104.49	37.39	$74,200,000,000,000

Table 3:

Variables	Coefficient	P-value	H0 test
Exports Goods & Services	-0.419	0.001431	Reject
Imports Goods & Services	-0.235	0.073715	Accept
GDP Annual Growth	0.587	0.000002	Reject
Urban Population Growth	0.697	0.000000	Reject

Table 4:

Factors Label	Beta	Factor Name
X121_2012	-0.16591	EXPORTS OF GOODS AND SERVICES (% OF GDP)
X134_2012	0	FOREST AREA (% OF LAND AREA)
X139_2012	0	GDP AT MARKET PRICES (CURRENT US$)
X140_2012	1.36521	GDP GROWTH (ANNUAL %)
X142_2012	0	GDP PER CAPITA (CURRENT US$)
X143_2012	0	GDP PER CAPITA GROWTH (ANNUAL %)
X154_2012	0	IMPORTS OF GOODS AND SERVICES (% OF GDP)
X156_2012	0	IMPROVED WATER SOURCE (% OF POPULATION WITH ACCESS)
X161_2012	0	INDUSTRY, VALUE ADDED (% OF GDP)
X162_2012	3.287812	INFLATION, CONSUMER PRICES (ANNUAL %)
X171_2012	0	LIFE EXPECTANCY AT BIRTH, FEMALE (YEARS)
X174_2012	0	LIFETIME RISK OF MATERNAL DEATH (%)
X179_2012	0	MANUFACTURING, VALUE ADDED (% OF GDP)
X195_2012	0	NET MIGRATION
X220_2012	0	POPULATION GROWTH (ANNUAL %)
X258_2012	0	RURAL POPULATION (% OF TOTAL POPULATION)
X283_2012	0	URBAN POPULATION (% OF TOTAL)
X284_2012	5.864563	URBAN POPULATION GROWTH (ANNUAL %)
X31_2012	0	AGRICULTURAL LAND (% OF LAND AREA)
X35_2012	6.83061	AGRICULTURE, VALUE ADDED (% OF GDP)
X36_2012	0	AGRICULTURE, VALUE ADDED (ANNUAL % GROWTH)
X45_2012	0	ARABLE LAND (% OF LAND AREA)

X9_2012	0	ADJUSTED NET NATIONAL INCOME (CURRENT US$)

Table 5:

cluster	index	X121_2012	X140_2012	X162_2012	X284_2012	X35_2012
0	98.8	-0.176782	0.287073	-0.173317	0.069637	0.064485
1	126.333333	0.138874	-0.989686	-0.415107	-0.796854	-0.697461
2	111.875	-0.628858	1.022273	1.091042	1.521951	1.317118

Table 6:

cluster	Mean of X129	SD of X129
0	119.04	8.955167
1	92.98	8.261593
2	139.43	26.712605

Table 7:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              X129_2012   R-squared:                       0.632
Model:                            OLS   Adj. R-squared:                  0.611
Method:                 Least Squares   F-statistic:                     30.06
Date:                Sun, 06 Dec 2020   Prob (F-statistic):           2.52e-08
Time:                        13:44:35   Log-Likelihood:                -153.22
No. Observations:                  38   AIC:                             312.4
Df Residuals:                      35   BIC:                             317.4
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept        119.0367      3.671     32.430      0.000     111.585     126.488
C(cluster)[T.1]  -26.0613      5.191     -5.021      0.000     -36.599     -15.523
C(cluster)[T.2]   20.3946      6.224      3.277      0.002       7.760      33.029
==============================================================================
Omnibus:                       10.808   Durbin-Watson:                   1.489
Prob(Omnibus):                  0.004   Jarque-Bera (JB):               28.967
Skew:                          -0.161   Prob(JB):                     5.13e-07
Kurtosis:                       7.265   Cond. No.                         3.59
==============================================================================
```

Table 8:

```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower    upper   reject
----------------------------------------------------
     0      1 -26.0613  0.001 -38.7631 -13.3596   True
     0      2  20.3946 0.0065   5.1657  35.6235   True
     1      2  46.4559  0.001   31.227  61.6848   True
----------------------------------------------------
```

Figure 1:

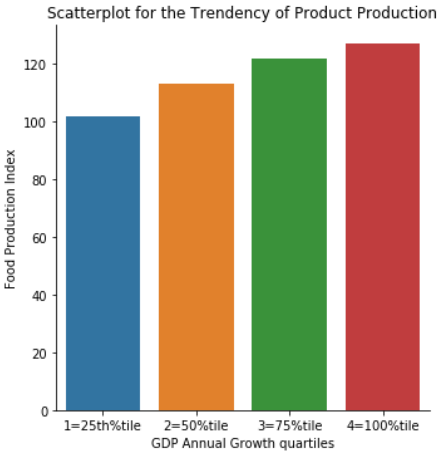


Figure 2:

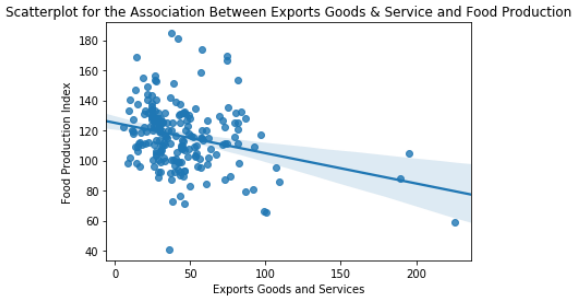


Figure 3:

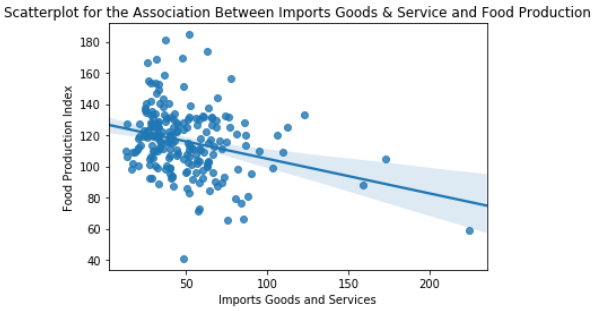


Figure 4:

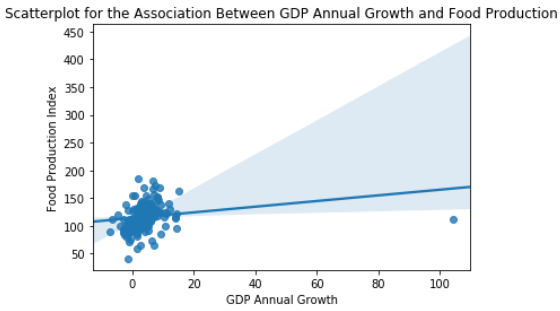


Figure 5:

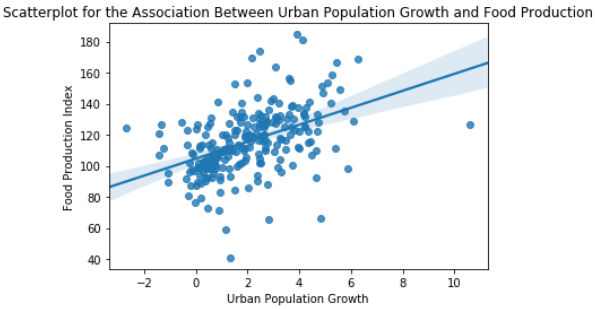


Figure 6:

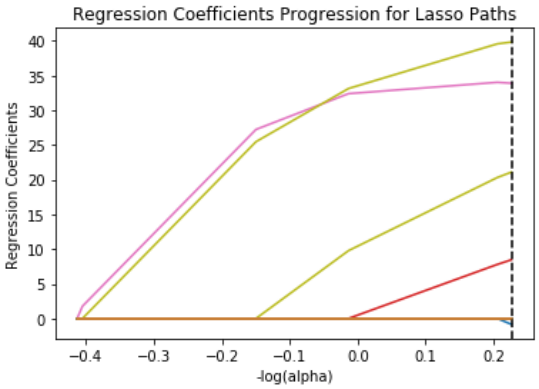


Figure 7:

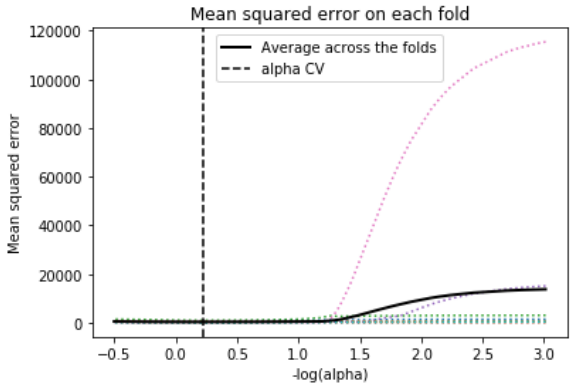


Figure 8:

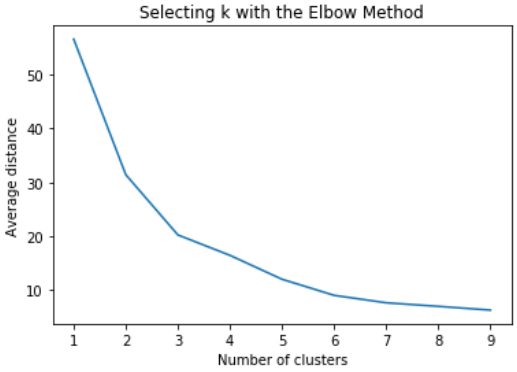


Figure 9:

