Impact of Canadian Crop Prices on Crop Production

Foundations of Data Science - Group Project

Muskaan Brar[1], Jeush Stanly[2], Monika Tomar[3], Jason Wilcox[4], Elmira Sepehri[5], and David Fishman[6]

Email ids: [1]mbrar44@uwo.ca, [2]jeush.stanly@mail.utoronto.ca, [3]25monikatomar25@gmail.com, [4]jason.wqost@slmails, [5]elmira.sphr@gmail, [6]davjfish@gmail.com

December 2023

Contents

1 Summary

1.1 Data source

1.1.1 What is the nature of the data you chose?

The dataset we first selected was sourced from the Government of Canada's Open Government Portal. It records the monthly price of various crop, commodity and livestock products produced in Canada. The raw data contains the price of each farm product, the month the price was captured, a product description, and the province of production. Our desire to perform analysis regarding the relationship between price and production led us to seek additional datasets containing production data for selected commodities. This production data was also sourced form Government of Canada's Open Government Portal. It recorded the monthly production of farm products by province of production over a period of years.

1.1.2 Why is it interesting?

The datasets we eventually selected were interesting as between them they contain a time series of prices for different products and from different provinces, and the production for those products over of a period of over time, from January 1980 to the present. The duration of the dataset provided an opportunity to find patterns and the possibility of a relationship between price and the amount of production within the regions of Canada. Analysis of these trends could provide us with insights with current day problems such as inflation and the cost of living crisis.

1.2 Analysis

1.2.1 What were the challenges?

Within our chosen datasets of Canadian farm product prices and production spanning from 1980 to present, our analysis required us to overcome a series of challenges that demanded a number of solutions.

The format of the certain features within the data, such as the REF_DATE were recorded as text and as such were not immediately usable, leading us to perform some feature engineering.

The structure of the pricing dataset allowed for differing units of measured to utilized within the data for a singled commodity.

The Wheat production dataset also could contain multiple entries for any given year which made comparison to other commodities problematic.

The differing ranges of data of the farm products represented in the dataset posed a challenge in terms of selecting the beat products for analysis

The publication of distinct the price and production datasets further added layers of complexity. Our analysis required these be merged.

The presence of null values in the merged pricing and production datasets added also required additional action prior to our final analysis.

Finally, the limited overlap in time series, resulted in a comparatively small sample size, and led to the selection of suitable techniques suitable for smaller datasets.

1.2.2 How did you overcome them?

To overcome the challenges presented in previous section, we took a number of steps to prepare the data and create a model for analysis

The date feature REF_DATE as imported as a text and feature engineering was leveraged to change its datatype to a date.

During the validation of the model, we noted that the Wheat production data reported production in March, July and December, as opposed once a year. As the March and July values were the same the previously reported December value, we were able to make the assumption that the December values where the correct ones. Accordingly, we selected the the December values and re-indexed the data to match the price data prior to merging.

The source of the data noted that the unit of measure many not be consistent. As part of our data cleaning we validated that the unit of measure for our the selected products was indeed consistent.

When reviewing a visual representing of the time series data for all provinces. It became evident that all the province had a similar patten, with the exception of Newfoundland and Labrador. Therefore we elected to exclude Newfoundland and Labrador from our subsequent analysis.

To analyze to impact of product production on product price it was needed to add production data to our model. Using the Government of Canada's Open Government Portal we search and found production data for our selected product products. After similar feature engineering to correct the REF_DATE feature, we merged the production data using the product description as a key.

The availability of price and production data led us to select Meat Chickens, Wheat and Hogs as products for final analysis on relationship between price and production. The availability of price and production data also required us to drop null values resulting in few data points for comparison.

To evaluate the relationship between price and production we applied the technique of linear regression on the resulting data model.

1.3 Conclusions

1.3.1 What did you learn about your dataset?

After reviewing the liner regression results for Meat Chickens, Wheat and Hogs we found the following:

For Meat Chickens, the model intercept was found to be 26.4810 and the coefficient price variable is 50.9459.

production = 26.4810 + 50.9459 x price

The F-staticic for this model was 12.47 and the changes of observing this statistic under a normal distribution is less than 0.05%. Therefore, with a p-value set to 0.05, we would reject the null hypothesis.

An R2 value of 0.675 means approximately 67% of the variance of the data can be accounted for by this model. We lead us to the conclusion that price does affect production in Meat Chickens and supports the regression of the null hypothesis.

The model for Wheat had an intercept of 1.285e4 and the coefficient for the price variable is 41.3883

production = 1.285e4 + 41.3883 x price

The F-staticic for this Wheat was 8.065 and the changes of observing this statistic under a normal distribution is less than 0.05%. Therefore, with a p-value set to 0.05, we would reject the null hypothesis.

The R2 is 0.28 which means approximately 30% of the variance of the data can be accounted for by this model. There is more unexplained variability in these data compared to the Meat Chickens

Finally the model for Hogs has an F-staticic of 0.3441 and the chances of observing this statistic under a normal distribution is approximately 56%. Therefore, with a p-value set to 0.05, we are not in a position to reject the null hypothesis.

Based on these results, we can say that only price affects production only in some farm products produced in Canada. This combined with some understanding of of Canadian economic policy, show that affect to be stronger in products without any form of Supply Management involvement from Government.

2 Objectives

2.1 What are the goals of the analysis and why did you choose them?

The goal of this analysis is to explore and better understand changes in the production of agricultural crops in Canada. Understanding these fluctuations can yield valuable insights into the Canadian economy and the agricultural sector.

Statistics Canada collects copious amounts of data via the Census of Agriculture (REF) thus making these datasets excellent candidates for analysis in our group project.

2.2 What question(s) do you want to answer?

While the collections of elements affecting the total production of agricultural products are complex and multifaceted, this report we focus solely on a single variables: price.

2.3 What hypothesis(es) do you have and what is your approach to tackle the problem?

The null hypothesis for this anal

NOTES: supply management, commodities, large investments

3 Data Preparation

3.1 What was your data source (e.g., web scraping, corporate data, a standard machine learning data set, open data, etc.)?

The dataset used in this study is sourced from open data published by Statistics Canada, a government institution. The use of open data from a reputable government source enhances transparency and allows for reproducibility in research

3.2 How good was the data quality?

3.3 What did you need to do to procure it?

To procure the dataset, we downloaded the csv of the dataset and we read it using the pandas library.

3.4 What tools or code did you need to use to prepare it for analysis?

First, we segmented our analysis into three distinct categories: chicken, hogs, and wheat. The main libraries we used for our analysis were pandas, seaborn, numpy and statsmodels.api.Utilizing a dedicated helper function, we crafted time series plots for each category, with a focus on prices—a pivotal variable in our analysis. To ensure data completeness, we applied a boolean mask to handle null values. While exploring the data we recognized the historical prices of Newfoundland is very different from other provinces as you can see in Figure 1 so we decided to exclude Newfoundland from our analysis. We decided that it would be easiest to work with the data if all the prices were aggregated into a single response so we created a time series for each category where the mean prices for each category is the data and the datetime objects as the index.

3.5 What challenges did you face?

One notable hurdle emerged with certain features, like 'ref date', initially recorded as text, necessitating a round of feature engineering to render them immediately usable. The pricing dataset, with its varied units of measurement for a single commodity, introduced complexity, prompting us to carefully address this diversity during our analysis. The Wheat production

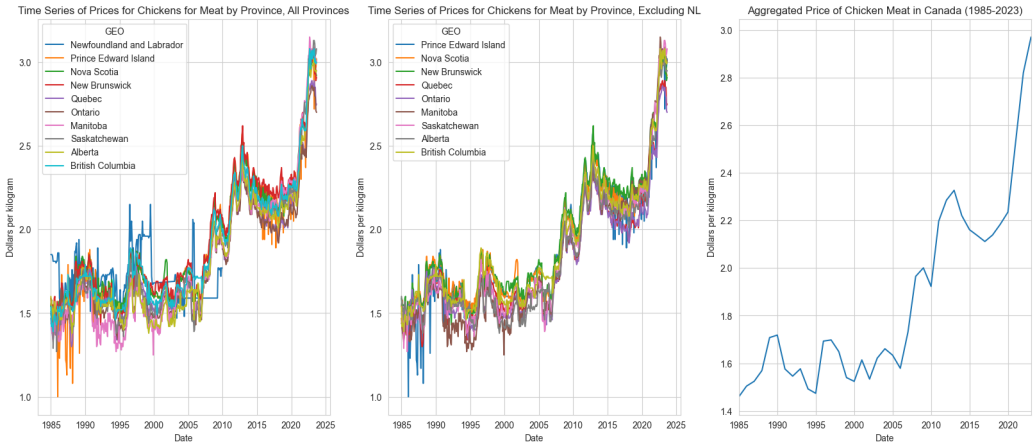

Figure 1: This figure shows the raw data, and aggregated forms of chicken prices. MORE TEXT NEEDED?

dataset presented its own intricacies, featuring multiple entries for a given year. This intricacy made comparisons with other commodities a nuanced task. Meanwhile, the diverse data ranges of farm products posed a challenge in selecting the most pertinent ones for our analytical lens. The separation of price and production datasets added layers of complexity, urging us to integrate these disparate sources for a more comprehensive understanding. Addressing null values within the merged datasets became another crucial step in ensuring the integrity of our analysis.

4 Analysis

4.1 What trends, correlations, and/or patterns do you see in the data?

Different trends were observed for different data. This section will outline the results of our analysis for the three crops.

4.1.1 Poultry

The data frame used for the final analysis contained eight observations. When the prices and production were plotted using a scatterplot, a clear linear relationship was observed. The scatterplot and trend-line for this relationship can be observed in Figure 2. Using the statsmodels package in Python, we determined the coefficients for the linear model using the ordinary least squares (OLS) approach. The output from the regression can be viewed in Figure 3.

The null hypothesis for this test was that no relationship exists between price and production. Following the regression, there was the resulting model:

$$y = 26.4810 + 50.9459x_1$$

- Where y is the price of chicken meat, in dollars per kilogram
- and x_1 is the number of chickens, measured in millions of individuals

The F-statistic for this model was observed to be 12.47; under a normal distribution the chances of observing this value are 0.01%. Setting our p-value to 0.05, we would be forced to reject the null hypothesis and conclude there is indeed a relationship between the price and production of chicken in Canada. Finally, The observed value for R^2 was 0.675 which means that approximately 67% of the variance of the data can be accounted for by this model.

4.1.2 Hog

4.1.3 Wheat

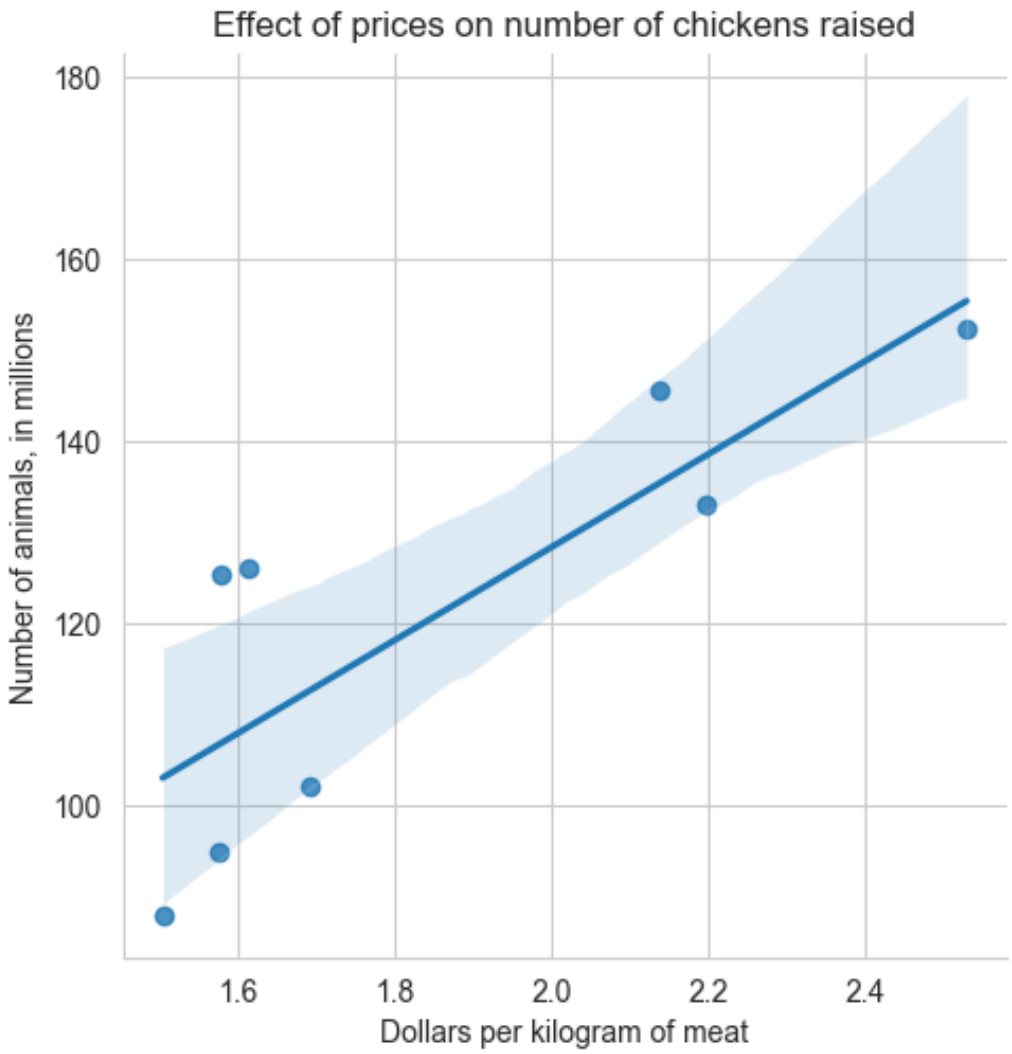

Figure 2: A scatter plot of chicken prices vs. chicken production.

OLS Regression Results

Dep. Variable:	production	**R-squared:**	0.675
Model:	OLS	**Adj. R-squared:**	0.621
Method:	Least Squares	**F-statistic:**	12.47
Date:	Sun, 10 Dec 2023	**Prob (F-statistic):**	0.0124
Time:	16:19:44	**Log-Likelihood:**	-31.609
No. Observations:	8	**AIC:**	67.22
Df Residuals:	6	**BIC:**	67.38
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
prices	50.9459	14.428	3.531	0.012	15.642	86.250
Eins	26.4810	27.235	0.972	0.368	-40.161	93.123

Omnibus:	2.302	**Durbin-Watson:**	1.404
Prob(Omnibus):	0.316	**Jarque-Bera (JB):**	0.929
Skew:	0.382	**Prob(JB):**	0.629
Kurtosis:	1.516	**Cond. No.**	12.7

Figure 3: Results from an Ordinary Least Squares Regression.

5 Conclusions

5.1 What did you learn about your data set?

Scratch:

This analysis is complex because it is not clear which variable should be the response and which are the covariates. For example, we can imagine a scenario where the increase of prices results in the increase of production, but it is also conceivable that the various levels of production will drive prices changes.