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 over time 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 insight to 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 January 1980 to present, our analysis required us to overcome a series of challenges.

The format of the certain features within the data, such as the REF_DATE were imported as a text datatype 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 be utilized within the data for a singled commodity. Also, the Wheat Production dataset could contain multiple entries for any given year which merging with the pricing data problematic. The differing ranges of data of for each of the farm products in the dataset posed a challenge in terms of selecting the best data for analysis The publication of distinct within different files for 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 to create and validate 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 December values and re-indexed the data to match the Price dataset 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 line plot 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 searched for 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 any 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.

$$y = 26.4810 + 50.9459x_1$$

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

The F-statistic for this model was observed to be 12.47; under a normal distribution the chances of observing this value are approximately 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 meat in Canada. 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.

Similarly, we found there to be a significant, positive relationship between the prices and production levels of wheat. The model for this relationship can be described as follows:

$$y = 1.285e4 + 41.3883x_1$$

- Where y is the production of wheat, in thousands of metric tonnes
- and x_1 is the price of wheat, in dollars per metric tonne

The F-statistic for this model was observed to be 8.065; under a normal distribution the chances of observing this value are approximately 0.01%. With a p-value set 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 wheat in Canada, albeit not as strong of a relationship as with chicken meat. The observed value for R^2 was 0.287 which means that approximately 29% of the variance of the data can be accounted for by this model. While this finding is significant, it is clear there are more factors involved in understanding the variations of production of wheat in Canada than just price alone.

Finally, we did not find there to be a significant relationship between the production of hogs and prices in Canada.

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 [1] 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. Specifically, we seek to answer the question of what affect does price have on the production of different agricultural products?

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

Our hypothesis is that there exists a relationship between the price of a commodity and its production in Canada, particularly for commodities without any Supply Management interference from the Government. To tackle the problem, we sourced data from the Government's Open Data Portal, prepared the data and then created and validated a model of select products. We then applied linear regression on that dataset to search for a relationship between price and production.

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 the Government of Canada, on its Open Government Portal [3], an official government website. 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?

In general, the quality of the data was very good. Aside from the minimal transformations and cleaning required, the dataset was structured in a way that facilitated our analysis. Specifically, each row of data related to a single observation and each column was an attribute of that datum. Occasionally, observations had overlapping data, for example, one row containing values for a single province and another containing data for all of Canada. When this occurred, it was important care was taken not to double-count those data.

3.3 What did you need to do to procure it?

To procure the dataset, we downloaded the csv of the dataset and we ingested it using the pandas library [2].

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 Meat, 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 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.

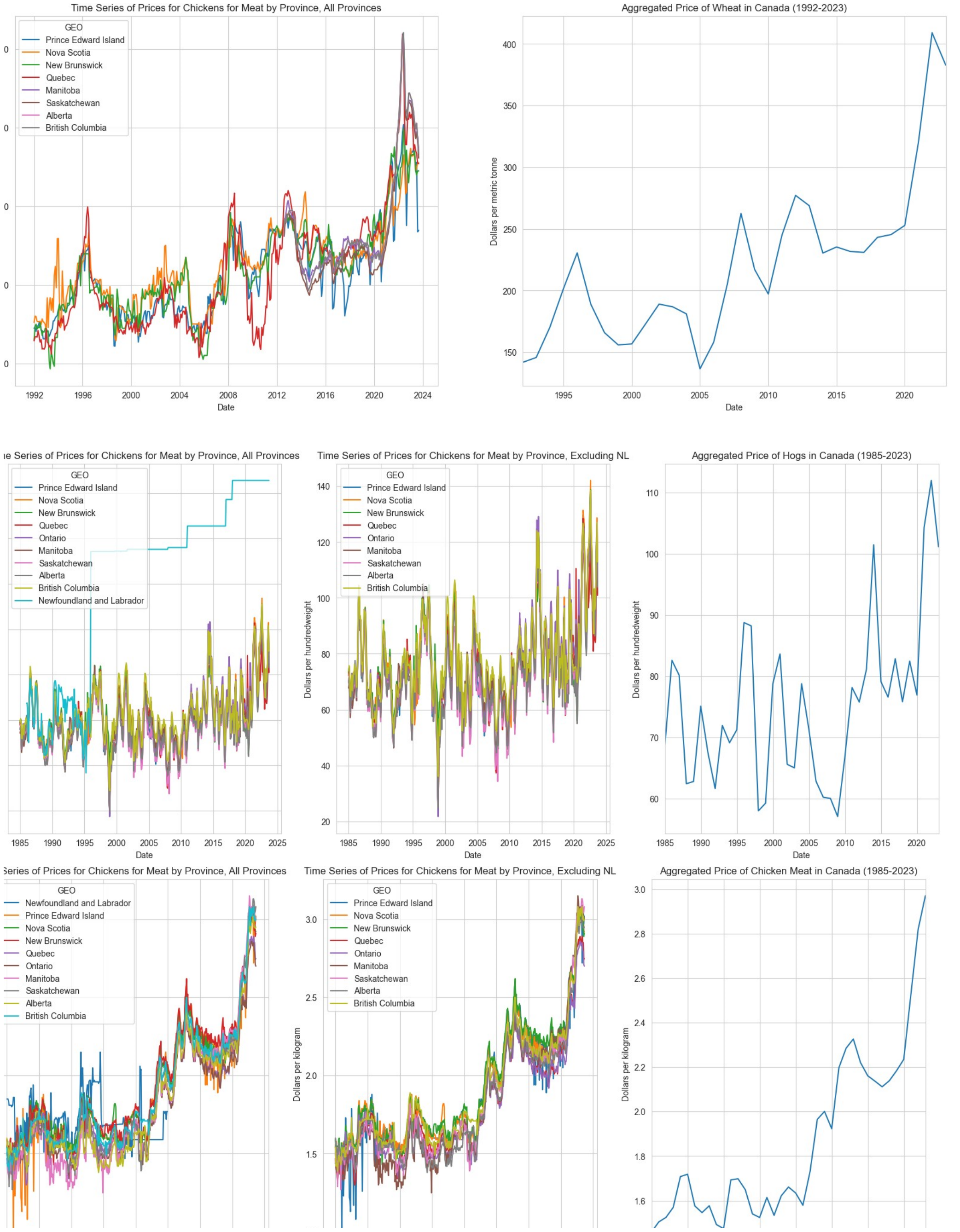


Figure 1: This figure shows the raw data, and aggregated forms of Chicken Meat, Hogs and Wheat.

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 Meat Chickens

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

The F-statistic for this model was observed to be 12.47; under a normal distribution the chances of observing this value are approximately 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 Hogs

The data frame used for the final analysis contained a total of 38 observations. When the prices and production were plotted using a scatterplot, the ellipsoid did not appear to have any meaningful structure. The scatterplot and trend-line are displayed in Figure 4. Using the statsmodels package in Python, we determined the coefficients for the linear model using the ordinary least

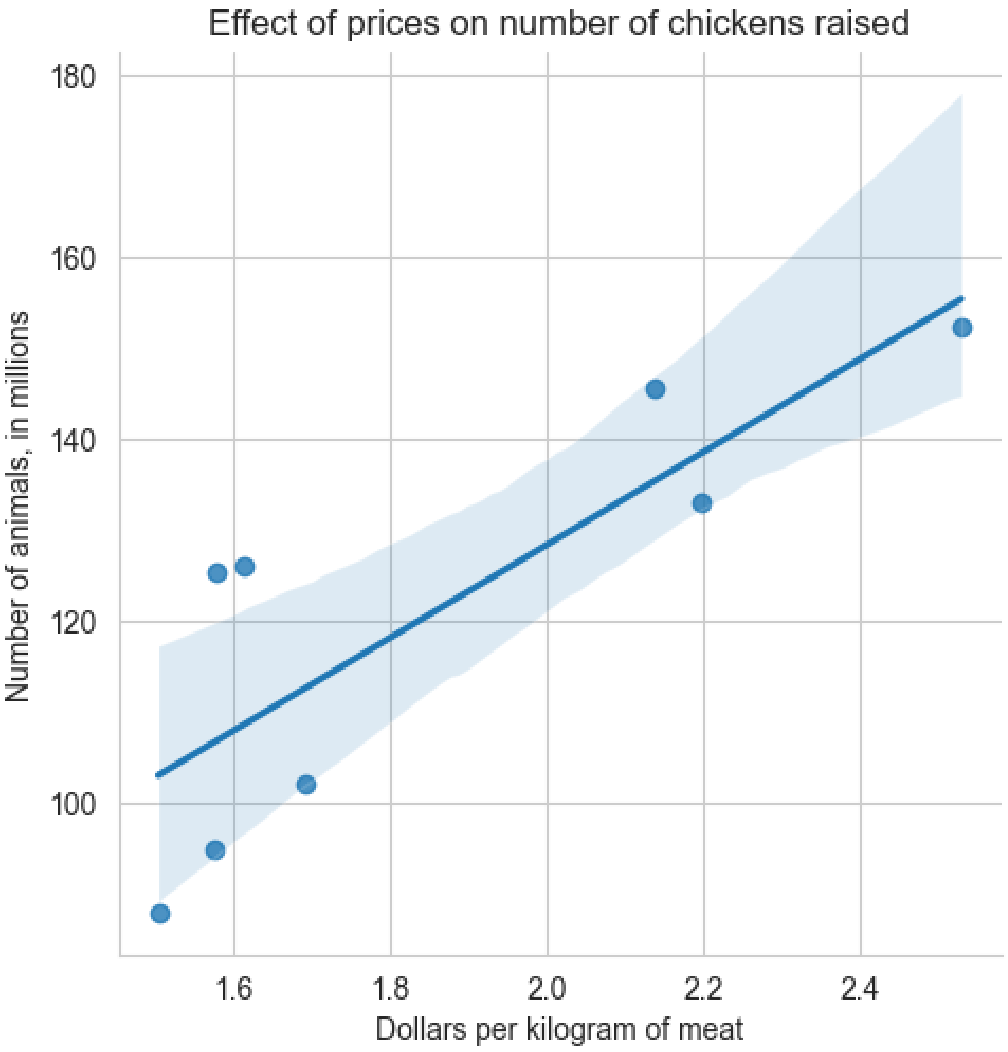
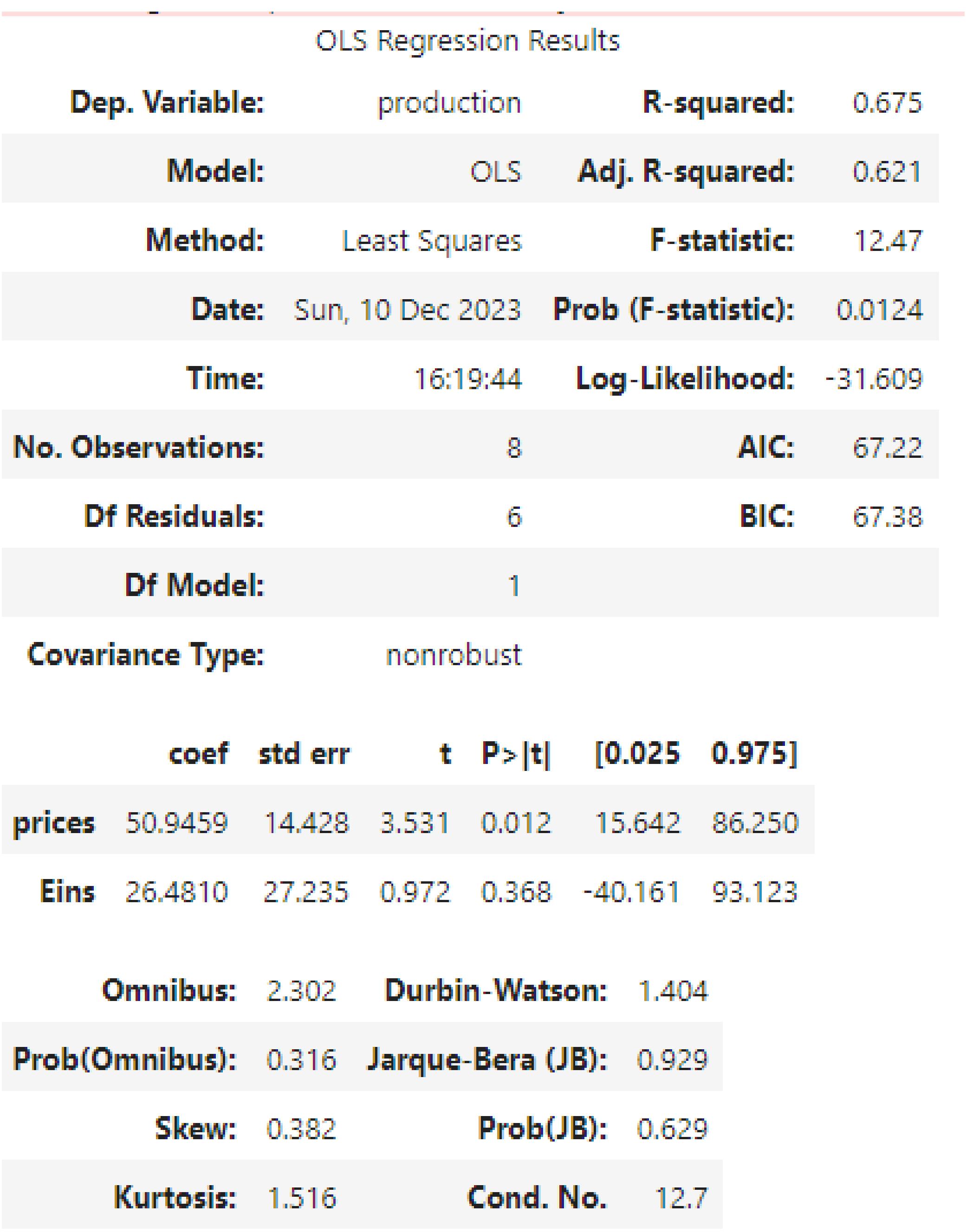


Figure 2: A scatter plot of Meat Chicken prices vs. Meat Chicken production.



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 performed on Chickens raised in Canada.

squares (OLS) approach. The output from the regression can be viewed in Figure 5.

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

$$y = 20.3703 + 0.0441x_1$$

- Where y is the estimated output of farms, measured in millions of individuals
- and x_1 is the price of hog, in dollars per hundredweight

The F-statistic for this model was observed to be 0.3441. With our p-value set to 0.05, the null hypothesis that no relationship exists between the price and production of hogs in Canada would be accepted. Based on the observed R^2 was 0.009, we can state that effective none of the variance in the dataset was explained by this model.

4.1.3 Wheat

The data frame used for the final analysis contained 22 observations. When the prices and production were plotted using a scatterplot, some degree of linearity was observed. The scatterplot and trend-line for this relationship can be observed in Figure 6. 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 7.

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

$$y = 1.285e4 + 41.3883x_1$$

- Where y is the production of wheat, in thousands of metric tonnes
- and x_1 is the price of wheat, in dollars per metric tonne

The F-statistic for this model was observed to be 8.065; under a normal distribution the chances of observing this value are approximately 0.01%. With a p-value set to 0.05, we would be forced to reject the null hypothesis and conclude there is indeed a relationship between the price and production

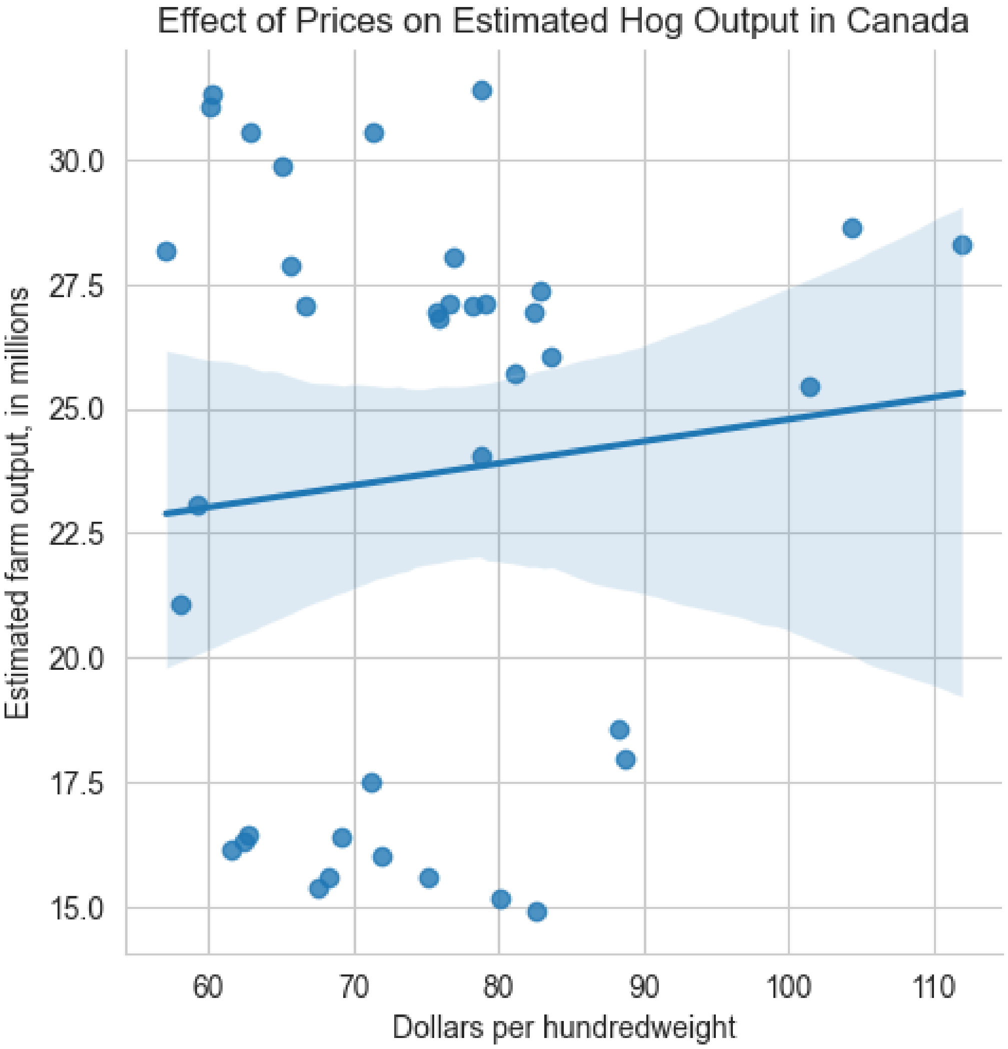
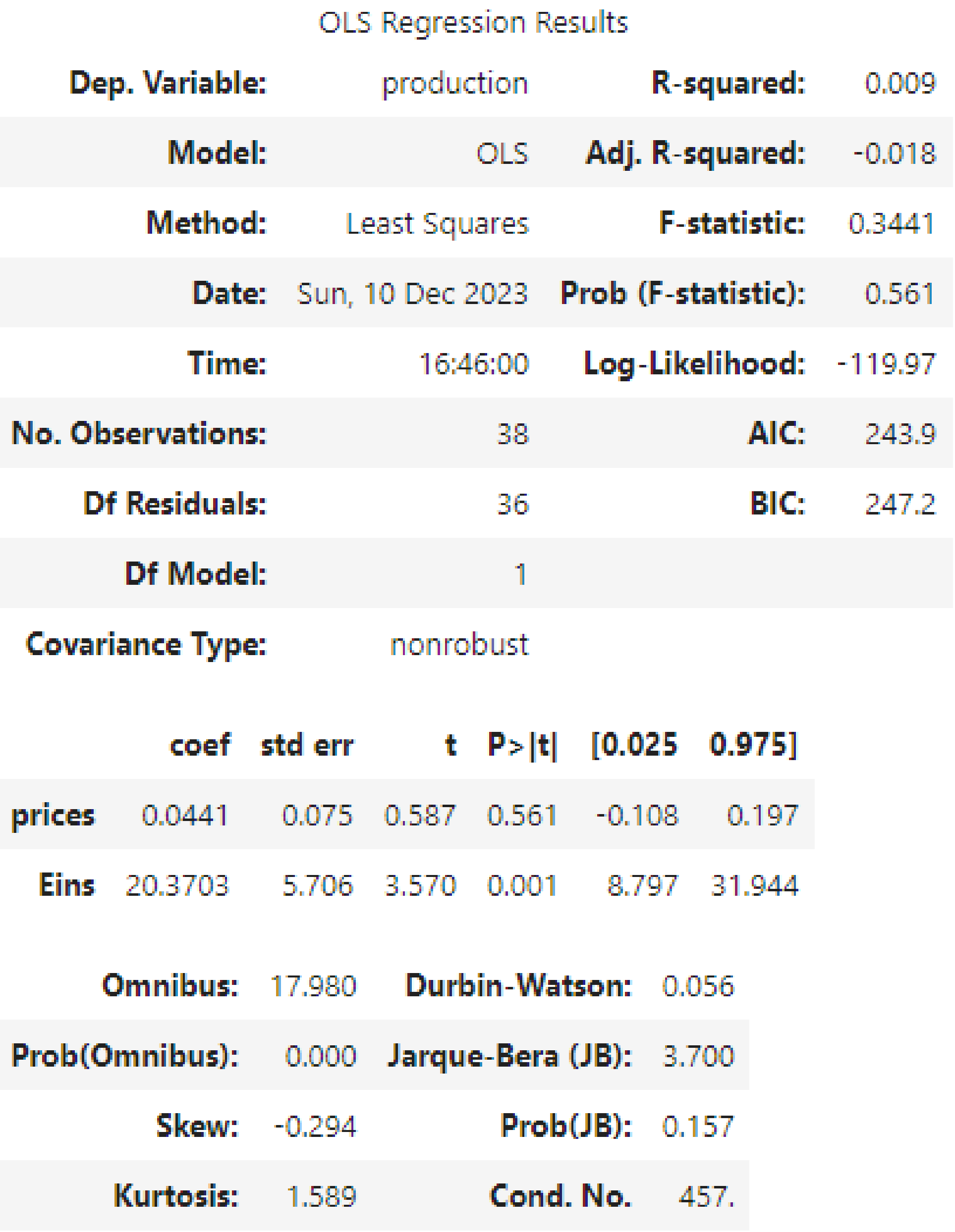


Figure 4: A scatter plot of hog prices vs. hog production in Canada.



OLS Regression Results

Dep. Variable:	production	**R-squared:**	0.009
Model:	OLS	**Adj. R-squared:**	-0.018
Method:	Least Squares	**F-statistic:**	0.3441
Date:	Sun, 10 Dec 2023	**Prob (F-statistic):**	0.561
Time:	16:46:00	**Log-Likelihood:**	-119.97
No. Observations:	38	**AIC:**	243.9
Df Residuals:	36	**BIC:**	247.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
prices	0.0441	0.075	0.587	0.561	-0.108	0.197
Eins	20.3703	5.706	3.570	0.001	8.797	31.944

Omnibus:	17.980	**Durbin-Watson:**	0.056
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3.700
Skew:	-0.294	**Prob(JB):**	0.157
Kurtosis:	1.589	**Cond. No.**	457.

Figure 5: Results from an Ordinary Least Squares Regression performed on hog production in Canada.

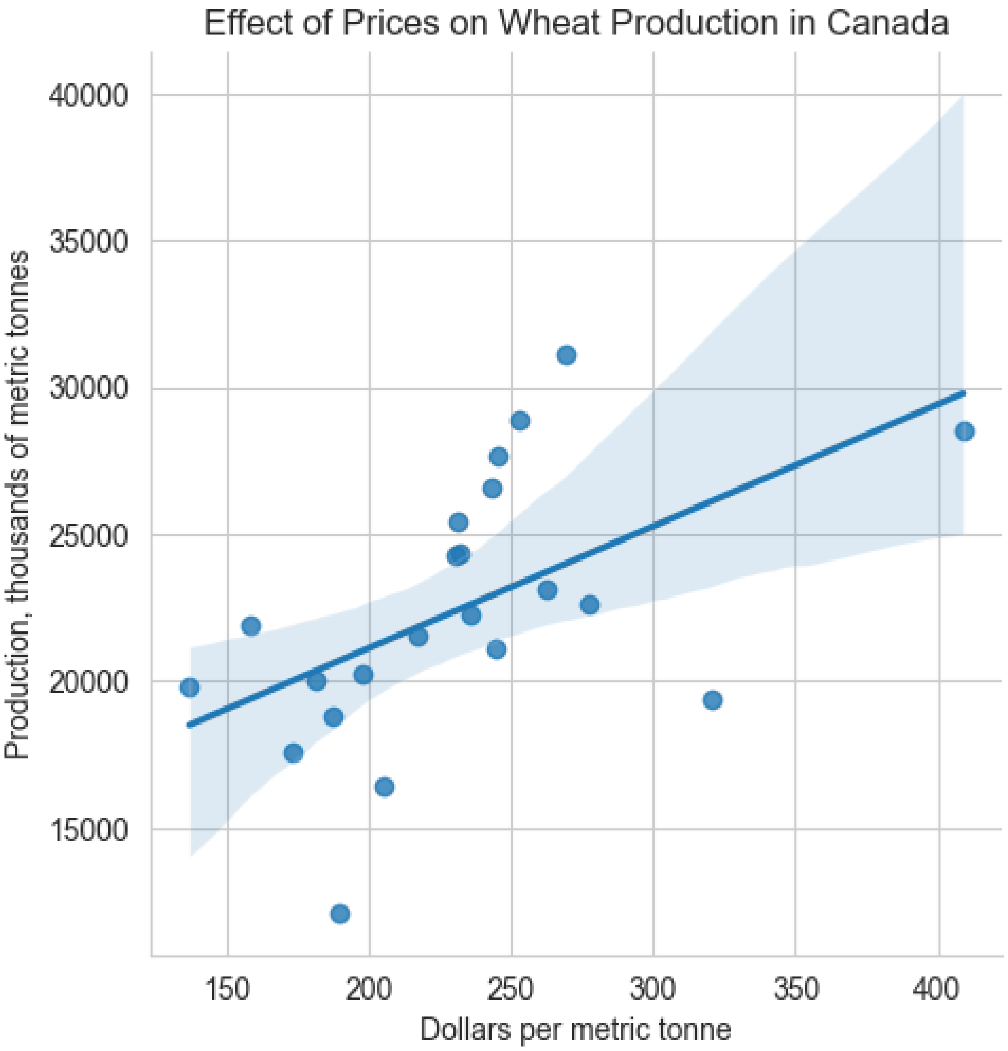
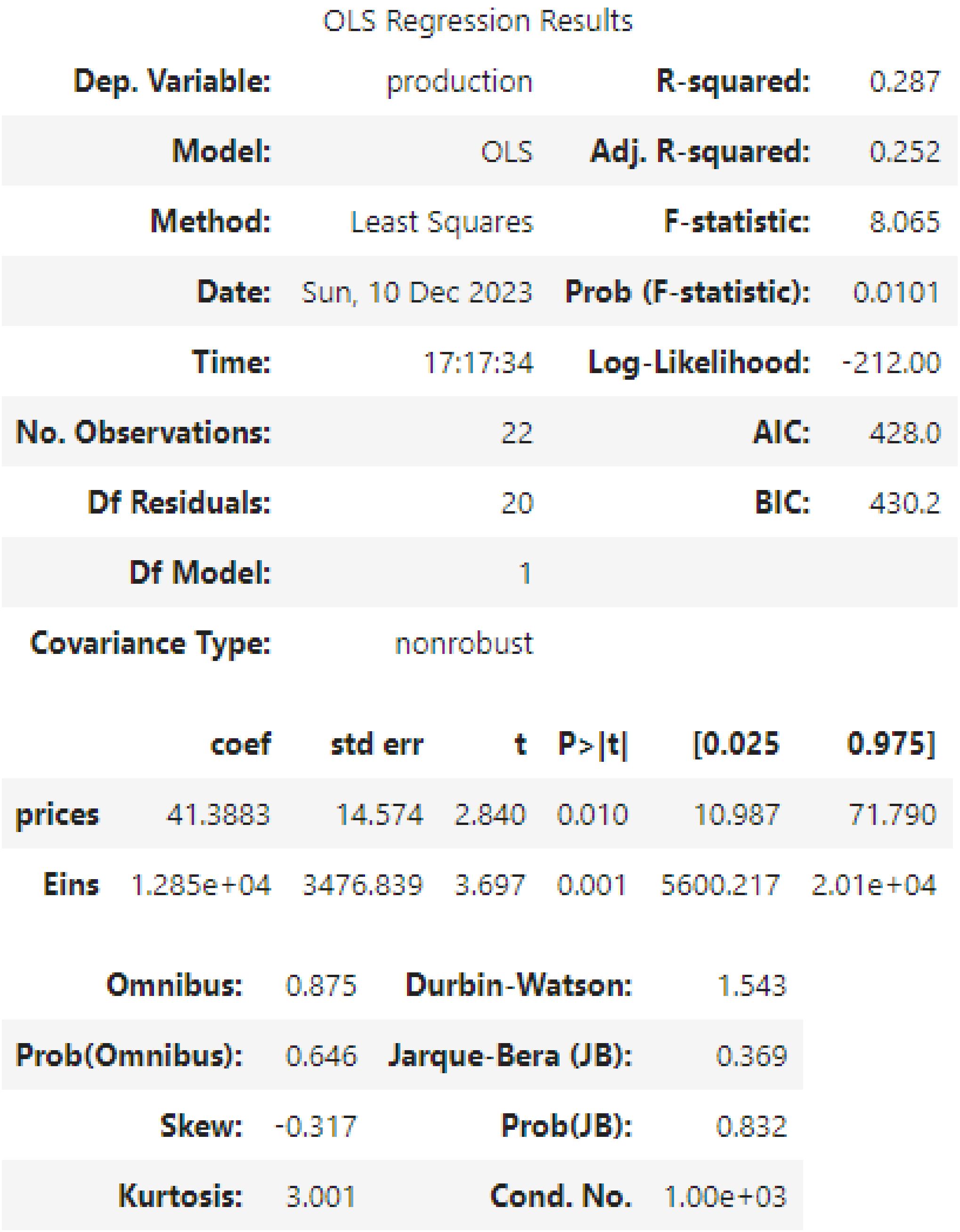


Figure 6: A scatter plot of wheat prices vs. wheat production.



OLS Regression Results

Dep. Variable:	production	**R-squared:**	0.287
Model:	OLS	**Adj. R-squared:**	0.252
Method:	Least Squares	**F-statistic:**	8.065
Date:	Sun, 10 Dec 2023	**Prob (F-statistic):**	0.0101
Time:	17:17:34	**Log-Likelihood:**	-212.00
No. Observations:	22	**AIC:**	428.0
Df Residuals:	20	**BIC:**	430.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
prices	41.3883	14.574	2.840	0.010	10.987	71.790
Eins	1.285e+04	3476.839	3.697	0.001	5600.217	2.01e+04

Omnibus:	0.875	**Durbin-Watson:**	1.543
Prob(Omnibus):	0.646	**Jarque-Bera (JB):**	0.369
Skew:	-0.317	**Prob(JB):**	0.832
Kurtosis:	3.001	**Cond. No.**	1.00e+03

Figure 7: Results from an Ordinary Least Squares Regression performed on wheat production in Canada.

of wheat in Canada, albeit not as strong of a relationship as with chicken meat. Finally, The observed value for R^2 was 0.287 which means that approximately 29% of the variance of the data can be accounted for by this model. While this finding is significant, it is clear there are more factors involved in understanding the variations of production of wheat in Canada than solely price.

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.

References

[1] Statistics Canada. Census of agriculture. https://www.statcan.gc.ca/en/census-agriculture, Dec 2021. Accessed on 2023-12-11.

[2] NumFOCUS. pandas. https://pandas.pydata.org/pandas-docs/stable/index.html, 2023. Accessed on 2023-12-11.

[3] Government of Canada. Open government portal. https://open.canada.ca/, 2023. Accessed on 2023-12-11.