Zhaocheng (Raymond) Gu
zg127@georgetown.edu·(240)885-9263·https://github.com/GGGGUKIM/Writing-sample

CODING SAMPLE

The following document is one of the coding tests that I completed. This test involves using Stata to clean, wrangle, and visualize data, and build econometric models. The last part also involves using econometric knowledge to answer several short questions. The answers are my own work. Please do not distribute it without my permission. The folder can also be accessed on my GitHub.

2. Data Cleaning

The central task of this section is to merge production data and price data into one dataset. Both data are sourced from the US Department of Agriculture. Both datasets contain annual data from 1990 to 2018 A brief introduction on the datasets:

- Barley production.csv lists the barley production in bushels by agricultural district. The agricultural district is an administrative division between the county and state levels.
- Barley price.csv lists the mean price received by farmers per bushel of barley by state. Ignore the distinction between the marketing year and the calendar year.

  We want the final dataset to be:
  • a panel with three dimensions: year, agricultural district, state • in each row it contains: barley production and price

3.1

For each year, compute the weighted average of price over all states, where each state's weight is its production in bushels in that year. Then, plot this weighted average over the time period from 1990 to 2018.

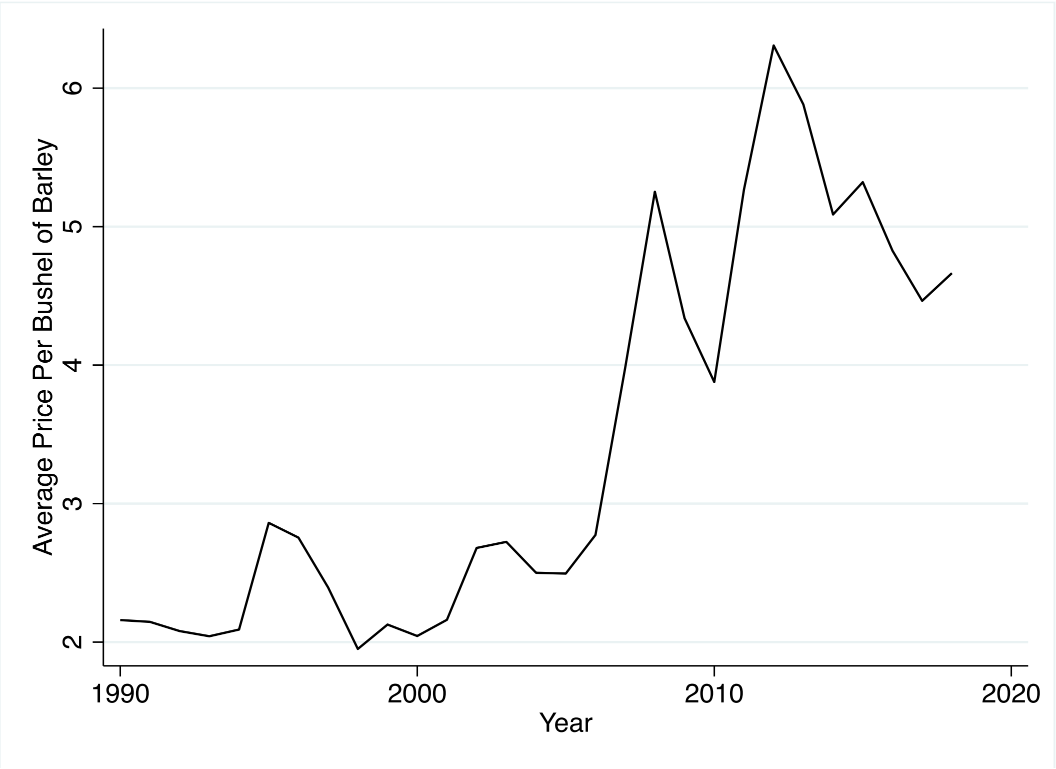


Figure 1. Time Series Plot: Price

3.2

Find the top 3 states in terms of barley production in 2018. Plot the time series of production for these 3 states in the same plot, over the period from 1990 to 2018. Scale the production variable so that it is in millions of bushels.

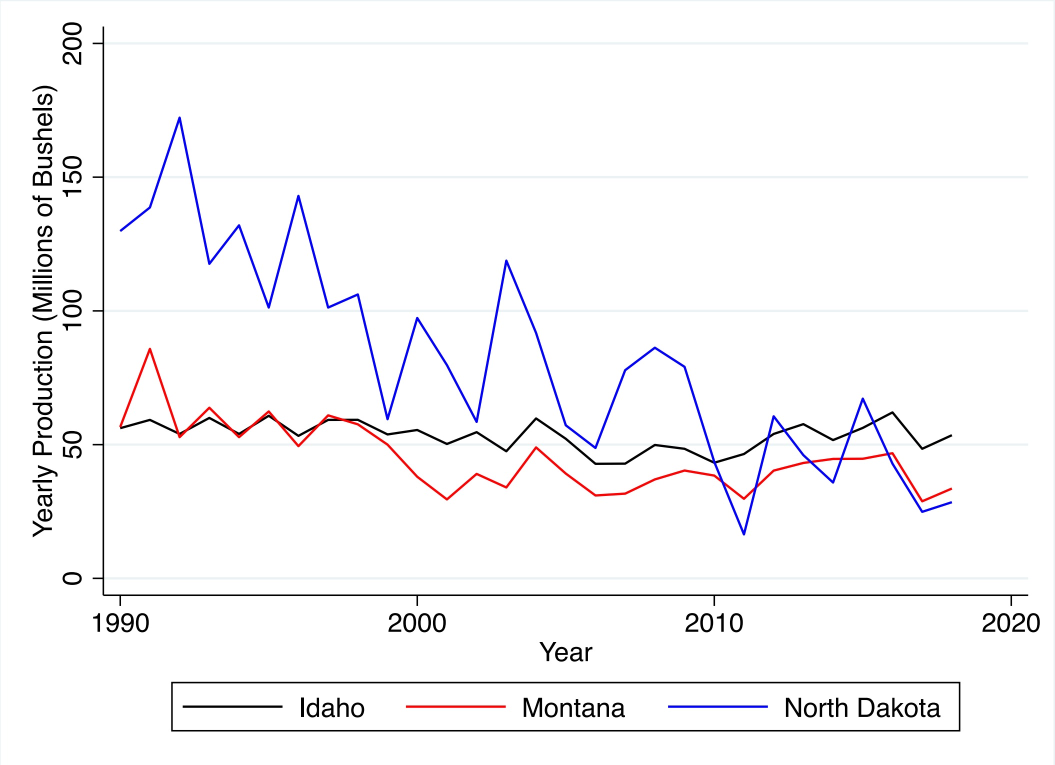


Figure 2. Time Series Plot: Production

3.3

Create a summary table where the rows are specific states (Idaho, Minnesota, Montana, North Dakota, and Wyoming) and the columns are decades (1990-1999, 2000-2009, and 2010-2018). The elements of the table are mean annual state-level production, by decade and state. Scale the production variable so that it is in millions of bushels.

Table 1. Summary Table (Millions of Bushels)

State	Production_1990_1999	Production_2000_2009	Production_2010_2018
IDAHO	56.99	50.4	52.6
MINNESOTA	32.95	7.73	5.19
MONTANA	59.22	36.88	38.92
NORTH DAKOTA	120.15	79.53	40.66
WYOMING	8.86	6.21	6.62

4. Short Answer

Our goal is to estimate the sensitivity of US farmers' barley production to barley price, using the provided data, at the level of agricultural district by year.

4.1

- First write down a regression equation of a linear model of production on a constant and price. We want the coefficient on price to have the interpretation of an elasticity. Ensure that the terms are properly indexed. Report the results of this regression, and interpret the coefficient on price.

Table 2. Results of the regression of lg_product on lg_price

	(1)
	lg_product
lg_price	1.224***
	(.083)
_cons	11.758***
	(.083)
Observations	3405
R-squared	.06

Standard errors are in parentheses

**** p<.01, ** p<.05, * p<.1*

1) From table 2, the regression uses 3405 observations and has an R-squared with .06. The constant is 11.758 and statistically significant at the conventional level. The coefficient on lg_price turns out to be statistically significant at the conventional level, and its magnitude is 1.224.
2) The coefficient can be interpreted as a one-percent increase in the mean price of barley is correlated with about a 1.22% increase in barley production.

4.2

- What variables do you think we should control for? Choose two and explain why they might help us identify the coefficient on price. These variables need not to be in the original dataset.

1) I think barley consumption (demand) and technology are two controls that needed to be included in the regression.
2) First, according to the price theory in economics, demand can be positively related to price. Also, high demand may stimulate farmers to grow more barley. Thus, if barley consumption is omitted from the regression, it can cause an upward bias to the coefficient on price.
3) Second, if technology advances, relevant processing, transportation, and storing costs can decrease as well as the price. And production can increase with more advanced technology. Therefore, if technology is omitted from the regression, it can cause a downward bias to the coefficient on price. Including these two controls in the model can help address omitted variable bias to some extent.

4.3

- Price is an endogenous variable in our model. Provide examples of two different types of endogeneity that could bias our estimated coefficient on price.

1) The first one is omitted variable bias. As stated in the last question, there can be a series of other factors that can be related to both barley price and production, such as consumption, price of substitutes and complements, technology, import, and weather. Since it is a binary linear regression, these omitted variables can cause the correlation between price (independent variable of interest) and the error term and thus endogeneity.
2) Second, there can be possible simultaneity in the relationship between barley price and production. High prices can stimulate production, but production (along with demand) can also affect price according to the price theory in economics. And this simultaneity can cause endogeneity as well.

4.4

- We can somewhat mitigate this problem by including year and state fixed effects. Run this regression on the provided data and report the estimated coefficient on price, along with its standard error. Justify the method you used to adjust the standard error. Is this coefficient causally interpretable?

Table 3. Results of regression accounting for state and year fixed effects

	(1)[1]
	lg_product
lg_price	-.464
	(.359)
_cons	13.999***
	(.374)
Observations	3405
R-squared	.563

Standard errors are in parentheses

*** $p<.01$, ** $p<.05$, * $p<.1$

1) The coefficient on price in this regression is about -.464 with a clustered standard error of about .359.
2) Since 28 state dummies are included in the regression, I have adjusted the standard error by clustering on 28 different states to correct for heteroskedasticity if the disturbances are not identically distributed over the states or there is within-state serial correlation in the idiosyncratic errors.
3) The coefficient on lg_price is statistically insignificant at the conventional level, and it seems that there is no correlation between barley price and production based on this dataset when accounting for state and year dummies. Furthermore, even if the coefficient was statistically significant at the conventional level, it would not be internally valid and cannot represent a causal relationship of price on production due to other types of endogeneity that including the year and state fixed effects cannot address, such as simultaneity, selection bias, measurement error, and remaining omitted variable bias.

[1] I ignore the coefficients on those state and year dummies since they are not our variables of interest.

4.5

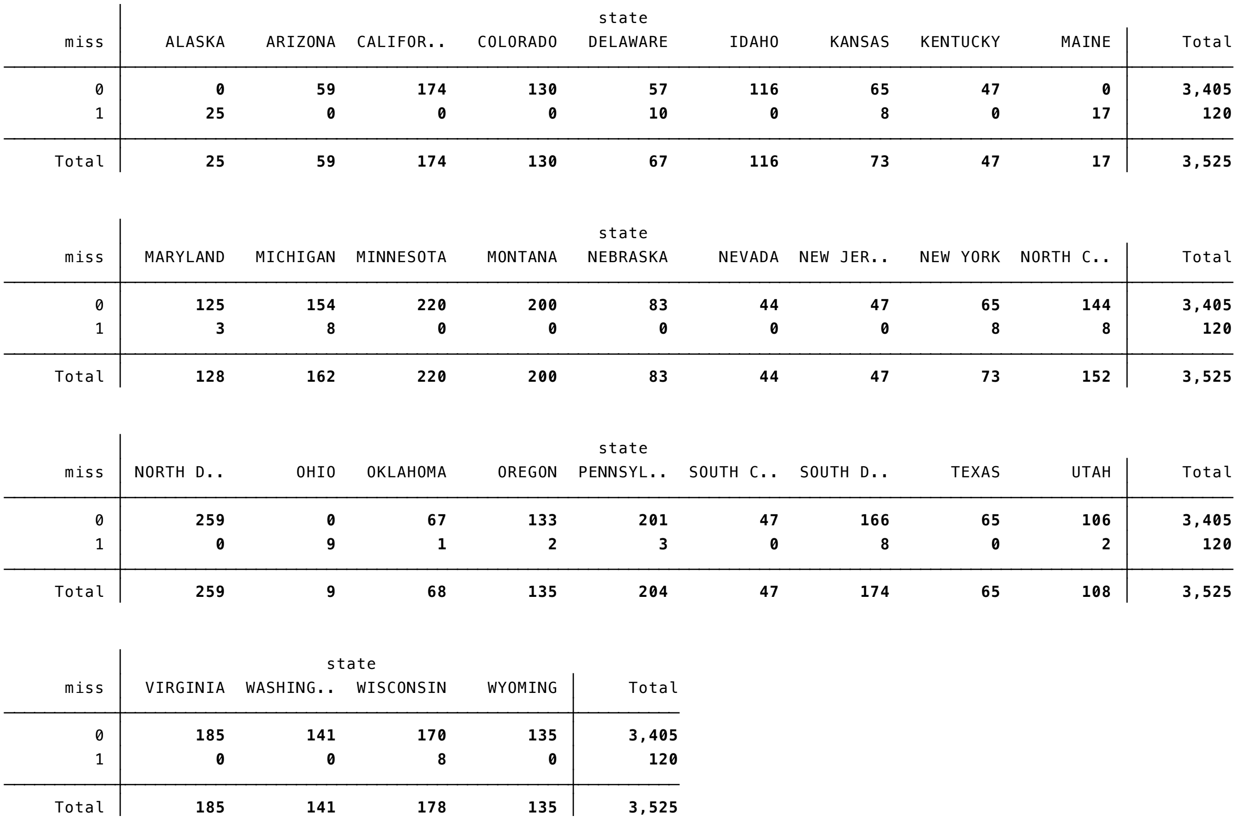
- Which potential sources of price endogeneity does adding fixed effects address? Discuss the difference (if any) in results with the results above. Which sources might still remain? Make sure to provide concrete examples.

1) Adding fixed effects helps address omitted variable bias to some extent, because these fixed effects control for all state-varying characteristics that are constant over time and all time-varying characteristics that are constant across states.
2) While the coefficient in the 4.1 is positive and statistically significant at the conventional level, the coefficient becomes negative and statistically insignificant at the conventional level after accounting for fixed effects. The constant and its standard error both become larger in the second regression compared to the first one, but the constant is still statistically significant at the conventional level. The number of observations used in the model does not change and the R-squared increases a lot.
3) In this regression, omitted variable bias, simultaneity, selection bias, and measurement error may still remain. First, although controlling for state and year fixed effects, factors that vary over time and across states such as import, technology, price of substitutes and complements, and weather are included in the error term, which can still cause the correlation between price and the error term and thus endogeneity. Second, despite controlling for only state and year dummies, it is still possible that production can affect price, and it can lead to endogeneity as well. Third, from Table 4 and Figure 3, we can find that there are incomparable characteristics between observations with missing observations and ones without, such as price and geographical location. This means the missing values may not occur at random, and can cause selection bias since Stata automatically omits observations with any missing value when doing regressions. Forth, some measurement errors may occur during the process of collecting data, which can cause endogeneity. For example, local farmers might misreport their barley production or some price data in distant areas were unable to obtain, which can cause the difference between documented data and real data.

Table 4. Results of regression of price on missing status

	(1) price
miss	.796***
	(.112)
_cons	2.715***
	(.021)
Observations	3525
R-squared	.014

Standard errors are in parentheses
****p<.01, **p<.05, *p<.1*



```
           |                                        state
      miss |    ALASKA    ARIZONA  CALIFOR..   COLORADO   DELAWARE      IDAHO     KANSAS   KENTUCKY      MAINE |     Total
-----------+---------------------------------------------------------------------------------------------------+----------
         0 |         0         59        174        130         57        116         65         47          0 |     3,405 
         1 |        25          0          0          0         10          0          8          0         17 |       120 
-----------+---------------------------------------------------------------------------------------------------+----------
     Total |        25         59        174        130         67        116         73         47         17 |     3,525 

           |                                        state
      miss |  MARYLAND   MICHIGAN  MINNESOTA    MONTANA   NEBRASKA     NEVADA  NEW JER..   NEW YORK  NORTH C.. |     Total
-----------+---------------------------------------------------------------------------------------------------+----------
         0 |       125        154        220        200         83         44         47         65        144 |     3,405 
         1 |         3          8          0          0          0          0          0          8          8 |       120 
-----------+---------------------------------------------------------------------------------------------------+----------
     Total |       128        162        220        200         83         44         47         73        152 |     3,525 

           |                                        state
      miss | NORTH D..       OHIO   OKLAHOMA     OREGON  PENNSYL..  SOUTH C..  SOUTH D..      TEXAS       UTAH |     Total
-----------+---------------------------------------------------------------------------------------------------+----------
         0 |       259          0         67        133        201         47        166         65        106 |     3,405 
         1 |         0          9          1          2          3          0          8          0          2 |       120 
-----------+---------------------------------------------------------------------------------------------------+----------
     Total |       259          9         68        135        204         47        174         65        108 |     3,525 

           |                    state
      miss |  VIRGINIA  WASHING..  WISCONSIN    WYOMING |     Total
-----------+--------------------------------------------+----------
         0 |       185        141        170        135 |     3,405 
         1 |         0          0          8          0 |       120 
-----------+--------------------------------------------+----------
     Total |       185        141        178        135 |     3,525 
```

Figure 3. Numbers of missing observations for each state

4.6

- To address any remaining sources of price endogeneity, suppose that we use the trans- portation cost between location of barley production and its market as an instrument for price. Explain why you think this is or is not a valid instrument. What challenges could arise when using this instrument in practice?

1) I think the transportation cost is not a valid instrumental variable here. To consider an IV's validity, we need to consider its relevance and exogeneity. For relevance, the transportation cost can be positively related to the price since the transportation cost will be included in the price when selling the goods. But for exogeneity, the transportation cost is always affected by local infrastructure building such as roads and bridges. And agricultural production is also always related to infrastructures such as reservoirs and electricity. Therefore, it is possible that the transportation cost can be related to barley production through other factors other than price, and this IV may not be exogenous and valid.
2) If we decide to use the transportation cost as an IV, there can be three challenges. First, it can be hard to collect data for transportation costs. It seems to be a lack of official systematic data about the transportation cost between the location of barley production and its market, especially for these distant time periods and areas. And collecting these data privately is quite demanding and time-consuming. Second, we need a strong correlation

between the key independent variable and the IV to keep the estimation precise (rule of thumb: a t-statistics larger than 3 on the coefficient in the first stage regression). There are a series of other factors that can be related to barley price except for the transportation cost, and the transportation cost may be one of these “important” factors. Therefore, the transportation cost can be a “weak” IV here. Third, the exclusion condition cannot be directly assessed statistically. Given the concerns on the exogeneity of the transportation cost (e.g., it can be related to production through infrastructure), this IV’s exogeneity needs to be checked, which can be hard and may not be so convincing.

4.7

- Supposeoneoftheotherresearchassistantsaccidentallydeletes10%oftheobservations of the barley production variable. How would you expect dropping these observations to change the estimated coefficient on price and its standard error if the deletions are at random? What if the deletions are not at random?

1) Random deletions will not bias the estimated coefficient on price, and the coefficient can still reflect the relationship (or causality if potential endogeneity problems can be solved) between price and barley production since there are no significant differences between the remaining observations and randomly deleted ones. However, the standard error will become larger given a smaller sample size (larger root MSE and fewer variations on the X), and the estimation will become less precise.
2) If the deletions are not at random, since deleted observations are not comparable with these remaining ones, selection bias can be more severe and the estimated coefficient on price can be biased and internally invalid. Also, the standard error will become larger with a smaller sample size.