FRE 501
Module 2 Assignment, Q2 and Q3
Due: 11:59 pm on November 26, 2021

This assignment (Q2 and Q3 combined) uses the USDA World Agricultural Supply Demand Estimates (WASDE) database for corn with harvested acres, yield and total use selected for the period April, 2010 to December, 2020. The WASDE data is transformed to create three variables: (1) the change in the forecasted acres from the previous report; (2) the change in the forecasted yield from the previous report; and (3) the change in the forecasted use from the previous report. The three change variables are linked to the change in the corn futures price from the previous trading day. The goal is to identify if there is statistical link between changes in the WASDE forecasts, which are contained in a particular report, and the change in the price on the day the report is released to commodity traders. We expect the futures price to increase (decrease) if the forecast for acres or yield is revised downward (upward). We also expect the futures price to increase (decrease) if the total use, which reflects the demand side of the market, is revised upward (downward).

The WASDE data is somewhat complicated because the forecasts are for end of season outcomes, which is the late fall for the case of U.S. corn. However the forecasts are reported by marketing year (September to the following August) and there is a nine month period where forecasts are being made for both the current marketing year and the next marketing year. One of the columns in the WASDE data is "ProjEstFlag". The data associated with the "Proj." flag (i.e., "projected") corresponds to forecasts for the next marketing year. The first forecast is in May and the last forecast is November, 18 months later. To simplify the data cleaning, in this assignment you will work with data from the first forecast in May through to the last forecast before the transition in April.

Q2 of this assignment involves cleaning and organizing the data, and Q3 involves statistical analysis. You must choose to complete both Q2 and Q3 in either Excel or R. If you choose to work with Excel then you will download the raw data from the USDA website and begin the cleaning process from scratch. If you will work with R then you will start with a partially cleaned data set (stored as a .Rds file). For the keen R users you have the option to conduct the initial cleaning yourself. By the end of the data cleaning exercise the data is approximately the same for Excel users and for R users.

If you will complete your assignment in R then download from Canvas the zip file named "AssignQ2_3_R.zip". Inside this zip file is a project file which must be opened first before running any of the code. The zip file also contains the data and a R Markdown template (made by Krisha), which gives you guidance for completing Q2 and Q3. Similar to Module 1 Assignment Q1 you will submit an HTML or PDF file which was rendered from your completed Markdown template.

If you will complete your assignment in Excel, then you will download the WASDE data from the USDA website, and you will download the corn futures price data from Canvas. You will answer Q2 and Q3 in the same Excel workbook, and you will answer the assignment discussion questions in a text box within the workbook. You will submit the completed workbook for grading.

The detailed instructions for completing the assignment in Excel are provided below.

Question 2: Data Cleaning

1. Download the zip file from https://www.usda.gov/oce/commodity/wasde titled "NEW: Consolidated Historical WASDE Report Data 2010- 2021" and unzip to reveal two .csv files. Open "oce-wasde-report-data-2010-04-to-2015-12.csv" in a new worksheet. Filter using Commodity = "Corn" and Attribute = c("Area Harvested", "Yield per Harvested Acre", "Use, Total") and ProjEstFlag = "Proj.". Uncheck "481" in the WasdeNumber filter – this report has some data errors. Copy and paste to a new worksheet with name "2010-2015". Close the original worksheet containing the full data. Save the workbook in Excel workbook format with name "MyWasdeData.xslx"
2. Open "oce-wasde-report-data-2016-01-to-2020-12.csv" and filter using Commodity = "Corn" and Attribute = c("Area Harvested", "Yield per Harvested Acre", "Use, Total"). Copy and paste to a new worksheet in "MyWasdeData.xslx" with name "2016-2020". Close the workbook containing the full data.
3. Use "Worksheet Copy" to copy sheet "2010-2015" to a new worksheet named "Combined". While omitting the header row, copy the contents of the "2016-2020" worksheet and paste immediately below the last row in the "Combined" worksheet.
4. Use "Worksheet Copy" to copy sheet "Combined" to a new worksheet named "Pivot". Delete all columns except "WasdeNumber", "Attribute" and "Value".

   The goal is to convert the WASDE data from long format to wide format. There is no direct way to do this in Excel but an indirect way is to use the Pivot Table tool. Aim to create the pivot table which appears in the image below.

Sum of Value	Column Labels			
Row Labels	Area Harvested	Use, Total	Yield per Harvested Acre	Grand Total
482	81.8	13300	163.5	13545.3
483	81.8	13410	163.5	13655.3
484	81	13360	163.5	13604.5
485	81	13490	165	13736
486	81	13440	162.5	13683.5

5. Copy the contents of the Pivot table (without the Grand Total column and row) to a new worksheet named "Wide". Rename the columns as follows:

	A	B	C	D
1	Wasde	Area	Use	Yield
2	482	81.8	13300	163.5
3	483	81.8	13410	163.5
4	484	81	13360	163.5
5	485	81	13490	165
6	486	81	13440	162.5

6. Sort the contents of "Wide" in increasing order of the "Wasde" column. The data should already be sorted but it is a good idea to use sort to ensure of the order.

7. Add two more column names to the right of “Wasde”. The first is “MktYear” (marketing year) and the second is “RelDate” (report release date). Use a lookup function and the “Combined” worksheet to insert the marketing year and report release date data in the new columns. Don’t worry about the duplicate values in the “WasdeNumber” column of “Combined” because for each WASDE number there is just one marketing year and one release date. The final data should look like as follows.

	A	B	C	D	E	F
1	Wasde	MktYear	RelDate	Area	Use	Yield
2	482	2010/11	5/11/2010	81.8	13300	163.5
3	483	2010/11	6/10/2010	81.8	13410	163.5
4	484	2010/11	7/9/2010	81	13360	163.5
5	485	2010/11	8/12/2010	81	13490	165
6	486	2010/11	9/10/2010	81	13440	162.5
7	487	2010/11	10/8/2010	81.3	13480	155.8
8	488	2010/11	11/9/2010	81.3	13430	154.3
9	489	2010/11	12/10/2010	81.3	13430	154.3

8. Download from Canvas the file “corn_price.csv”. Copy this worksheet to MyWasdeData.xlsx. Create a new column in the “corn_price” worksheet with name “P_Diff” and insert the lag price into this column (i.e., price in the cell one row higher).

9. Copy the “Wide” Worksheet and rename it to “Merged” (place at end of worksheets, to the right of “corn_price”. Delete the top row of data (i.e., Wasde 482). Rename the columns: “Area” to “Area_ch”, “Use” to “Use_ch” and “Yield” to “Yield_ch” (the “ch” refers to “change”). Referring to the “Wide” sheet, add formulas in the three renamed columns to calculate the difference between the current month value and the previous month value.

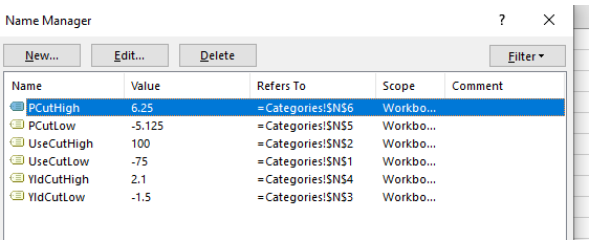
10. Add the name “P_Diff” to a new column in the “Merged” sheet. Use a lookup function with the “RelDate” column and “corn_price” sheet to add the corn price to the new column. Your results should look as follows.

	A	B	C	D	E	F	G
1	Wasde	MktYear	RelDate	Area_ch	Use_ch	Yield_ch	P_Diff
2	483	2010/11	6/10/2010	0	110	0	5
3	484	2010/11	7/9/2010	-0.8	-50	0	-2.25
4	485	2010/11	8/12/2010	0	130	1.5	10.75
5	486	2010/11	9/10/2010	0	-50	-2.5	7.75
6	487	2010/11	10/8/2010	0.3	40	-6.7	30
7	488	2010/11	11/9/2010	0	-50	-1.5	-9

11. Copy the “Merged” worksheet to a new worksheet named “Categories”. List the following six labels in a column: “UseCutLow”, “UseCutHigh”, “YldCutLow”, “YldCutHigh”, “PCutLow”, and “PCutHigh”. Use the Median and IF function in an array formula to calculate the median value of “Use_ch” for all negative values. Name this cell “UseCutLow”. Do the same for the positive values of “Use_ch” and assign to “UseCutHigh”. Repeat these steps for the “Yld_ch” and

"P_Diff" columns. The end result is as follows.

M	N
UseCutLow	-75
UseCutHigh	100
YldCutLow	-1.5
YldCutHigh	2.1
PCutLow	-5.125
PCutHigh	6.25



12. In the "Categories" worksheet create three new columns named "Use_Rank", "Yld_Rank" and "P_Rank". In the top empty cell of the "Use_Rank" column use a nested IF statement to generate the following text output.

→ "Use_Up_High" if the cell's value is greater than "UseCutHigh";
→ "Use_Up_Low" if the cell's value is greater than zero and less than "UseCutHigh";
→ "Use_NoChg" if the cell's value is equal to zero;
→ "Use_Down_Low" if the cell's value is greater than "UseCutLow" and less than zero; and
→ "Use_Down_High" if the cell's value is less than "UseCutLow".

Repeat this procedure for the "Yld_Rank" and "P_Rank" columns. The "Up" refers to an increase in the forecast from the month before, and "Down" refers to a decrease. The "NoChg" refers to no change in the variable from the month before. Your results should look as follows:

	A	B	C	D	E	F	G	H	I	J
1	Wasde	MktYear	RelDate	Area_ch	Use_ch	Yield_ch	P_Diff	Use_Rank	Yld_Rank	P_Rank
2	483	2010/11	6/10/2010	0	110	0	5	Use_Up_High	Yld_NoChg	P_Up_Low
3	484	2010/11	7/9/2010	-0.8	-50	0	-2.25	Use_Down_Low	Yld_NoChg	P_Down_Low
4	485	2010/11	8/12/2010	0	130	1.5	10.75	Use_Up_High	Yld_Up_Low	P_Up_High
5	486	2010/11	9/10/2010	0	-50	-2.5	7.75	Use_Down_Low	Yld_Down_Hig	P_Up_High
6	487	2010/11	10/8/2010	0.3	40	-6.7	30	Use_Up_Low	Yld_Down_Hig	P_Up_High

Question 3: Statistical Analysis of the Data

13. Copy the worksheet and name it "CrossTabs". Delete all columns except the first three and the three ranked columns. Important: Copy the values in the three ranked values and use Paste Special Values to convert the formulas to values. Your results should look as follows.

	A	B	C	D	E	F
1	Wasde	MktYear	RelDate	Use_Rank	Yld_Rank	P_Rank
2	483	2010/11	6/10/2010	Use_Up_High	Yld_NoChg	P_Up_Low
3	484	2010/11	7/9/2010	Use_Down_Low	Yld_NoChg	P_Down_Low
4	485	2010/11	8/12/2010	Use_Up_High	Yld_Up_Low	P_Up_High
5	486	2010/11	9/10/2010	Use_Down_Low	Yld_Down_Hig	P_Up_High
6	487	2010/11	10/8/2010	Use_Up_Low	Yld_Down_Hig	P_Up_High
7	488	2010/11	11/9/2010	Use_Down_Low	Yld_Down_Hig	P_Down_High

Values, not formulas

14. In the "CrossTabs" sheet use the Pivot Table tool to create cross tabulations for: (1) "Use_Rank" and "Yld_Rank"; (2) "Use_Rank" and "P_Rank"; and (3) "Yld_Rank" and "P_Rank". Your results

should look as follows.

H	I	J	K	L	M	N
Count of Yld_Ran	Column Labels					
Row Labels	Yld_Down_High	Yld_Down_Low	Yld_NoChg	Yld_Up_High	Yld_Up_Low	Grand Total
Use_Down_High	12	4	7	2	2	27
Use_Down_Low	3	6	12	1	2	24
Use_NoChg	1		19		2	22
Use_Up_High	2	2	7	10	2	23
Use_Up_Low	2	4	15	1	7	29
Grand Total	**20**	**16**	**60**	**14**	**15**	**125**
Count of P_Rank	Column Labels					
Row Labels	Use_Down_High	Use_Down_Low	Use_NoChg	Use_Up_High	Use_Up_Low	Grand Total
P_Down_High	7	5	6	6	9	33
P_Down_Low	4	10	7	4	8	33
P_NoChg	2		1			3
P_Up_High	8	2	3	5	8	26
P_Up_Low	6	7	5	8	4	30
Grand Total	**27**	**24**	**22**	**23**	**29**	**125**
Count of P_Rank	Column Labels					
Row Labels	Yld_Down_High	Yld_Down_Low	Yld_NoChg	Yld_Up_High	Yld_Up_Low	Grand Total
P_Down_High	4	3	9	10	7	33
P_Down_Low	3	3	24		3	33
P_NoChg	1		2			3
P_Up_High	9	5	10		2	26
P_Up_Low	3	5	15	4	3	30
Grand Total	**20**	**16**	**60**	**14**	**15**	**125**

15. Create a text box in a new worksheet names “Answers” and provide answers to the following questions.

    a. How common is it to have a “win-win” scenario for long traders where forecasted yield is revised strongly downward and forecasted use is forecasted upward? Provide a specific percentage.

    b. Would you characterize the unconditional distribution of use and yield as given by the row and column grand totals as being reasonably well approximated by a normal distribution, or is there a noticeable skew in the distribution?

    c. The top two rows of each pivot table correspond to small and large price decreases in response to the forecast. The left (right) two columns correspond to a forecasted decrease (increase) in the WASDE variable. In the middle pivot table corresponding to use, the four values in the top left are therefore “correctly” classified according to theory, and the four values in the top right are “incorrectly” classified. The opposite is true for the bottom pivot table, which correspond to forecasted yield. How many “correct” and “incorrect” classifications are there for use and for yield? Based on a comparison of these values, is it the change in forecasted use or yield which is more strongly associated with the change in price?

16. Your goal is to conduct a Chi Square test of association between the change in the WASDE use forecast and the change in the futures price. You will then conduct a second Chi Square test using the change in the WASDE yield forecast and the change in the futures price. Copy the "CrossTabs" sheet and name it "ChiSquare". To conduct the test follow the steps in the video https://www.youtube.com/watch?v=KNR59hgPC4E .

17. Answer the following questions in the text box of your "Answers" sheet.
    d. Are you able to reject the null hypothesis that the association between the change in the forecasted WASDE use and the change in the futures price is random; i.e., there is no statistical association? If yes, at what level of confidence?

    e. Are you able to reject the null hypothesis that the association between the change in the forecasted WASDE yield and the change in the futures price is random; i.e., there is no statistical association? If yes, at what level of confidence?

    f. Briefly discuss the disadvantage of using a Chi Square test which tests for the association between pairs of variables, versus regression which allows for a full array or explanatory variables. Hint: When working with quantitative data, think about the advantage of regression over simple correlation.

18. To address the shortcomings discussed in your answer (f) you will regress the price change variable (P_diff) on use and yield categorical variables. Copy the "Crosstabs" sheet and name it "Regression". Delete the pivot tables. Now create two new columns named "use_cat" and "yield_cat". Use a nested IF statement to assign categorical values:

    -2 for "Use_Down_High" and "Yld_Down_High",
    -1 for "Use_Down_Low" and "Yld_Down_Low",
    0 for "Use_NoChg" and "Yld_NoChg",
    1 for "Use_Up_Low" and "Yld_Up_Low" and
    2 for "Use_Up_High" and "Yld_Up_High".

19. Regress "P_diff" on "use_cat" and "yield_cat". Answer the following questions in the text box of the "Answer" sheet.
    g. Do the estimated coefficients have the anticipated sign, and are they statistically significant?
    h. What does the specific value of the coefficient on the "use_cat" variable represent (your answer should be stated in quantitative terms, similar to the coefficient of a dummy variable)?
    i. What does the specific value of the regression intercept represent?