Assignment 2: Q2 and Q3 Instructions and Code Guide

# Instructions

For questions 2 and 3 of Assignment 2, you will be working with the USDA World Agriculture Supply and Demand Estimates (WASDE) monthly reports from 2010-2020. You can read more about it [here](https://www.usda.gov/oce/commodity/wasde).

**Question 2** involves cleaning the WASDE and corn data (2 points). **Question 3** involves analyzing the effect of the WASDE report on corn prices. There are 2 models you can choose from to answer Question 3 (3 points).

1. **Monthly Regression** - Regress change in price from previous day on the change in forecast from previous month.
2. **Event study** - Use daily prices as the dependent variable and USDA report categorical variables as explanatory variables.

The raw data has been processed for you. We have imported 2 raw data files, filtered the corn reports and 3 attributes (acres, yield, and use) only, and saved it as wasde\_corn\_proj.rds file. You can also try to replicate this part yourself; hints are provided too!

Instructions on how to proceed with Questions 2 and 3 are provided. You can refer to the Q2\_codetips.pdf file on how to use the functions suggested in this assignment. You can also refer to the R bootcamp notes [here](https://raw.githack.com/limkrisha/mfre-r-bootcamp-2021/main/presentation/intro-to-r.html) and [here](https://raw.githack.com/limkrisha/mfre-r-bootcamp-2021/main/presentation/working-with-data.html) for additional reference. The pdf file should be sufficient; only readRDS(), left\_join(), lag(), and lm() functions are not included there, but most of you have used these functions in your Assignment 1. We also suggest you collaborate through Piazza, but remember to complete the assignment yourself.

**Expected Output**. You can fill in the codes in this Markdown file. You will have to un-comment the suggested code and fill in the correct code where it says insert\_code\_here. If you see codes that prints the output but are commented out, such as table(wasde$Attribute) or head(all\_data), kindly un-comment these lines, so I can see whether you are on the right track. Knit this file to either html or pdf. Submit the html or pdf file on Canvas.

pacman::p\_load(here, dplyr, ggplot2, janitor, tidyr)

## Processing the Raw Data - Optional exercise

Download historical USDA WASDE Report data from their [website](https://www.usda.gov/oce/commodity-markets/wasde/historical-wasde-report-data). Unzip the two folders: April 2010-December 2015 and January 2016 to December 2020. Copy and paste the csv files to your “Data” folder associated with this R Project.

* Using the read.csv() function, load the 2010-2015 report in csv format into R and call it dataFirst.
* Using the read.csv() function, load the 2016-2020 report in csv format into R and call it dataSecond.
* Use the rbind() function to (row)bind these two dataframes together and call it data. You will end up with 617,465 observations.
* Use the filter() function to filter observations where Commodity == Corn and ProjEstFlag == Proj. and call this new dataframe wasde. You will now have 17,200 observations.
* Using the saveRDS() and here() function, save this wasde dataframe as wasde\_corn\_proj.rds in your “Data” folder.

# load in the wasde files  
dataFirst <- read.csv(here("Data", "oce-wasde-report-data-2010-04-to-2015-12.csv"))  
dataSecond <- read.csv(here("Data", "oce-wasde-report-data-2016-01-to-2020-12.csv"))  
  
# combine these two files using the rbind() function (rbind = rowbind)  
data <- rbind(dataFirst, dataSecond)  
  
# create new dataframe called wasde that contains only the corn commodity from the data dataframe  
wasde <- filter(data, Commodity == "Corn", ProjEstFlag == "Proj.")  
  
# save this dataframe as an RDS file and call it wasde\_corn\_proj.RDS  
saveRDS(wasde, here("Data", "wasde\_corn\_proj.rds"))

Now it’s time for you to start coding!

# Question 2: Data Cleaning

* Using the readRDS() and here() functions, load the wasde\_corn\_proj.rds data from the Data folder and call it wasdeAll.
* Next, using the select() function, drop the following columns: ReportDate, ReportTitle, ReliabilityProjection, Region, AnnualQuarterFlag, ReleaseTime, Unit, ProjEstFlag. Call this dataframe wasde.
* Using the head() function, print the first 15 rows of the wasde() dataframe

wasdeAll <- readRDS(here("Data", "wasde\_corn\_proj.rds"))  
  
wasde <- select(wasdeAll, -c("ReportDate", "ReportTitle","ReliabilityProjection", "Region",  
"AnnualQuarterFlag","ReleaseTime", "Unit", "ProjEstFlag"))  
  
# print 15 rows of the wasde dataframe  
head(wasde, 15)

## WasdeNumber Attribute Commodity MarketYear Value ReleaseDate  
## 1 481 Area Harvested Corn 2009/10 79.6 2010-04-09  
## 2 481 Area Planted Corn 2009/10 86.5 2010-04-09  
## 3 481 Avg. farm price - High Corn 2009/10 3.7 2010-04-09  
## 4 481 Avg. farm price - Low Corn 2009/10 3.5 2010-04-09  
## 5 481 Beginning stocks Corn 2009/10 1673.0 2010-04-09  
## 6 481 CCC inventory Corn 2009/10 0.0 2010-04-09  
## 7 481 Domestic, total Corn 2009/10 11015.0 2010-04-09  
## 8 481 Ending stocks Corn 2009/10 1899.0 2010-04-09  
## 9 481 Ethanol for fuel Corn 2009/10 4300.0 2010-04-09  
## 10 481 Exports Corn 2009/10 1900.0 2010-04-09  
## 11 481 Feed and residual Corn 2009/10 5450.0 2010-04-09  
## 12 481 Food, seed & industrial Corn 2009/10 5565.0 2010-04-09  
## 13 481 Free stocks Corn 2009/10 1899.0 2010-04-09  
## 14 481 Imports Corn 2009/10 10.0 2010-04-09  
## 15 481 Outstanding loans Corn 2009/10 175.0 2010-04-09  
## ForecastYear ForecastMonth  
## 1 2010 4  
## 2 2010 4  
## 3 2010 4  
## 4 2010 4  
## 5 2010 4  
## 6 2010 4  
## 7 2010 4  
## 8 2010 4  
## 9 2010 4  
## 10 2010 4  
## 11 2010 4  
## 12 2010 4  
## 13 2010 4  
## 14 2010 4  
## 15 2010 4

Right now, the wasde dataframe is in a long format. For this analysis, we need to data to be in a wide format. If you take a look at the output of table(wasde$Attribute), you will notice that there are two different categories for “Use, Total” because of capitalization issues (i.e., most is “Use, Total” and one entry is "Use, total); the same too for some other variables. The differences in capitalization comes from the report WasdeNumber == 481.

For this particular exercise, we will just drop the first report (i.e., WasdeNumber == 481). Using the filter() function, we will filter observations for which WasdeNumber is not equal to 481.

wasde <- filter(wasde, WasdeNumber != 481)  
dim(wasde)

## [1] 17181 8

Some of the column names are long. Using the rename() function, rename

* ReleaseDate to Release
* ForecastYear to Forecast
* ForecastMonth to Month

wasde <- wasde %>% rename(Date = ReleaseDate,  
 Year = ForecastYear,  
 Month = ForecastMonth)

Then use filter() and the %in% or | operators to filter observations where Attribute takes the value of Area Harvested, Yield per Harvested Acre, and Use, Total only.

wasde <- wasde %>%  
 filter(Attribute %in% c("Area Harvested", "Yield per Harvested Acre", "Use, Total"))  
  
table(wasde$Attribute)

##   
## Area Harvested Use, Total Yield per Harvested Acre   
## 126 126 126

Now, we are ready to reshape the data (e.g., convert long to wide).

* use pivot\_wider() to reshape the data from long to wide format. Read [here](https://tidyr.tidyverse.org/reference/pivot_wider.html) for more info.
* use mutate() to convert Date to date format
* use rename() to rename Area Harvested to Acres, Use, Total to Use, and Yield per Harvested Acre to Yield

wasde\_wide <- wasde %>%  
 pivot\_wider(names\_from = Attribute,  
 values\_from = Value) %>%  
 mutate(Date = as.Date(Date, format = "%Y-%m-%d")) %>%  
 rename(Acres = `Area Harvested`,  
 Use = `Use, Total`,  
 Yield = `Yield per Harvested Acre`)

The wasde\_wide() data is now ready for analysis!

# Question 3: Analysis

Use the read.csv() and here() functions to load the corn\_price.csv file. Call this dataframe corn.

corn <- read.csv(here("Data", "corn\_price.csv"))  
head(corn)

## Date corn\_price  
## 1 1/1/2010 414.25  
## 2 1/4/2010 418.50  
## 3 1/5/2010 418.75  
## 4 1/6/2010 421.75  
## 5 1/7/2010 417.50  
## 6 1/8/2010 423.00

## Fill this section if you want to do Model 1 (Monthly Regression)

*Overview*. We want to analyze the effect of the change in USDA forecast in yield, acres, and use, respectively, from the *month before* on the change in price from the *day before*. To perform this analysis, we first calculate the change in corn prices from the day before () in the corn dataframe. Next, using the wasde\_wide dataframe, we calculate the change in yield, acres, and use, respectively, from the month; recall that the wasde\_wide dataframe contains monthly observations because the WASDE report is released monthly. Then we join these two dataframes together so that we can estimate how the monthly change in yield, acres, and use forecasts in the WASDE report affect the change in price from the day before. Hint: you should have 126 (monthly) observations.

Do the following transformations in the corn dataframe.

* Best to use %>% operator
* Use the rename() function to rename corn\_price to P\_current
* Use mutate() to convert Date column to a date format - check current format using head(Date) (hint: the date format in the corn dataframe is different from wasde dataframe)
* Use mutate() to create a new variable called P\_diff that calculates - (hint: Use lag() function, as in varname - lag(varname))

corn <- corn %>%  
 rename(P\_current = corn\_price) %>%  
 mutate(Date = as.Date(Date, format = c("%m/%d/%Y")),  
 P\_diff = P\_current - lag(P\_current))

Create a new dataframe called all\_data that contains a left join of wasde\_wide and corn dataframes, so that all rows of the wasde\_wide and only matching rows in the corn dataframe will be returned.

all\_data <- left\_join(wasde\_wide, corn, by = c("Date"))  
head(all\_data)

## # A tibble: 6 x 11  
## WasdeNumber Commodity MarketYear Date Year Month Acres Use Yield  
## <int> <chr> <chr> <date> <int> <int> <dbl> <dbl> <dbl>  
## 1 482 Corn 2010/11 2010-05-11 2010 5 81.8 13300 164.  
## 2 483 Corn 2010/11 2010-06-10 2010 6 81.8 13410 164.  
## 3 484 Corn 2010/11 2010-07-09 2010 7 81 13360 164.  
## 4 485 Corn 2010/11 2010-08-12 2010 8 81 13490 165   
## 5 486 Corn 2010/11 2010-09-10 2010 9 81 13440 162.  
## 6 487 Corn 2010/11 2010-10-08 2010 10 81.3 13480 156.  
## # ... with 2 more variables: P\_current <dbl>, P\_diff <dbl>

dim(all\_data)

## [1] 126 11

Do the following transformations in the all\_data dataframe.

* Best to use the %>% operator
* Use mutate() to create 3 variables A\_diff, U\_diff, and Y\_diff. Each variable takes the difference between and of Acres, Use, and Yield, respectively
* Use slice() to drop the first row

all\_data <- all\_data %>%  
 mutate(A\_diff = Acres - lag(Acres),  
 U\_diff = Use - lag(Use),  
 Y\_diff = Yield - lag(Yield)) %>%  
 slice(-1)

The marketing year changes from April to May, so we should not include the difference between April and May forecast in the analysis because these span two different marketing years. So now, you have to create a new dataframe called NoMay where you use filter() function to drop May observations.

noMay <- filter(all\_data, Month != 5)

Finally, you can run the model using the lm() function.

model\_lm <- lm(P\_diff ~ A\_diff + U\_diff + Y\_diff, data = noMay)  
summary(model\_lm)

##   
## Call:  
## lm(formula = P\_diff ~ A\_diff + U\_diff + Y\_diff, data = noMay)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.863 -5.359 -0.179 5.287 39.104   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.03313 0.96845 -0.034 0.973   
## A\_diff -0.57770 1.39428 -0.414 0.679   
## U\_diff 0.04356 0.01026 4.246 4.53e-05 \*\*\*  
## Y\_diff -2.44785 0.58782 -4.164 6.20e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.1 on 111 degrees of freedom  
## Multiple R-squared: 0.1535, Adjusted R-squared: 0.1306   
## F-statistic: 6.711 on 3 and 111 DF, p-value: 0.0003334

**Interpret your results**

* In 1-2 sentences, explain the intuition of the signs of the coefficients. Are they consistent with economic theory?
* Satellite data can now estimate acreage, so there is little information value in the USDA acreage forecast. However, these satellites are not able to estimate yield, and only USDA is able to estimate use. Is this story consistent with the findings in your regression? Why or why not?

# Fill in this section if you want to do Model 2 (event study)

**In your wasde\_wide dataframe, use the mutate() function to create 3 variables called d\_acres, d\_use, d\_yield.**

* Each variable can take only three values: 1, 0, -1 (hint: use nested ifelse() function)
* Variable = 1 if change from previous month > 0 (e.g., if )
* Variable = 0 if no change from previous month (e.g., if )
* Variable = -1 if change from previous month <0 (e.g., if )

wasde\_wide <- wasde\_wide %>%  
 mutate(d\_acres = ifelse(Acres - lag(Acres)>0, 1,  
 ifelse(Acres - lag(Acres)==0, 0,  
 ifelse(Acres-lag(Acres)<0, -1, NA))),  
 d\_use = ifelse(Use - lag(Use)>0, 1,  
 ifelse(Use - lag(Use)==0, 0,   
 ifelse(Use-lag(Use)<0, -1, NA))),  
 d\_yield = ifelse(Yield - lag(Yield)>0, 1,  
 ifelse(Yield - lag(Yield)==0, 0,   
 ifelse(Yield-lag(Yield)<0, -1, NA))))

**Create a new dataframe called corn\_wasde that contains a left join of corn and wasde\_wide dataframes, so that all rows of corn data frame and only matching rows in the wasde\_wide data frame will be returned.**

corn\_wasde <- left\_join(corn, wasde\_wide, by = c("Date"))  
head(corn\_wasde)

## Date P\_current P\_diff WasdeNumber Commodity MarketYear Year Month Acres  
## 1 2010-01-01 414.25 NA NA <NA> <NA> NA NA NA  
## 2 2010-01-04 418.50 4.25 NA <NA> <NA> NA NA NA  
## 3 2010-01-05 418.75 0.25 NA <NA> <NA> NA NA NA  
## 4 2010-01-06 421.75 3.00 NA <NA> <NA> NA NA NA  
## 5 2010-01-07 417.50 -4.25 NA <NA> <NA> NA NA NA  
## 6 2010-01-08 423.00 5.50 NA <NA> <NA> NA NA NA  
## Use Yield d\_acres d\_use d\_yield  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA

**Now you can run the event study model** **with the lm() function.**

eventstudy <- lm(P\_current ~ lag(P\_current) + d\_acres + d\_use + d\_yield, data = corn\_wasde)  
summary(eventstudy)

##   
## Call:  
## lm(formula = P\_current ~ lag(P\_current) + d\_acres + d\_use + d\_yield,   
## data = corn\_wasde)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -33.895 -5.265 0.642 6.146 33.712   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.308618 3.184321 0.097 0.9230   
## lag(P\_current) 0.998046 0.006584 151.584 < 2e-16 \*\*\*  
## d\_acres -0.610767 1.740374 -0.351 0.7262   
## d\_use 2.177205 1.215035 1.792 0.0757 .   
## d\_yield -6.707086 1.396961 -4.801 4.59e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.33 on 120 degrees of freedom  
## (2855 observations deleted due to missingness)  
## Multiple R-squared: 0.9949, Adjusted R-squared: 0.9948   
## F-statistic: 5881 on 4 and 120 DF, p-value: < 2.2e-16

**Interpret your results.**

* In 1-2 sentences, explain the intuition of the signs of the coefficients. Are they consistent with economic theory?
* Satellite data can now estimate acreage, so there is little information value in the USDA acreage forecast. However, these satellites are not able to estimate yield, and only USDA is able to estimate use. Is this story consistent with the findings in your regression? Why or why not?