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.

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).

a. **Monthly Regression** - Regress change in price from previous day on the change in forecast from previous month.
b. **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 and here 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. 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 <- insert_code_here
# dataSecond <- insert_code_here

# combine these two files using the rbind() function (rbind = rowbind)
# data <- rbind(insert_code_here)

# create new dataframe called wasde that contains only the corn commodity from the data dataframe
# wasde <- filter(insert_code_here)

# save this dataframe as an RDS file and call it wasde_corn_proj.RDS
# saveRDS(insert_code_here)
```

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 <- insert_code_here

# wasde <- insert_code_here

# print 15 rows of the wasde dataframe
# insert_code_here
```

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 <- insert_code_here
# dim(wasde)
```

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

- `ReleaseDate` to `Release`
- `ForecastYear` to `Forecast`

- `ForecastMonth` to `Month`

```
# wasde <- insert_code_here
```

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. You should only have 126 observations by now.

```
# wasde <- insert_code_here

# table(wasde$Attribute)
```

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 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 = insert_code_here,
#               values_from = insert_code_here) %>%
#   mutate(insert_code_here)) %>%
#   rename(insert_code_here)
```

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 <- insert_code_here
# head(corn)
```

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 ($P_t - P_{t-1}$) 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 P_t - P_{t-1} (hint: Use `lag()` function, as in `varname - lag(varname)`)

```
# corn <- insert_code_here
```

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. You should have 126 observations.

```
# all_data <- insert_code_here
# head(all_data)
# dim(all_data)
```

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 t and $t-1$ of `Acres`, `Use`, and `Yield`, respectively
- Use `slice()` to drop the first row

```
# all_data <- insert_code_here
```

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. Your `noMay` dataframe should have 115 observations.

```
# noMay <- insert_code_here
```

Finally, you can estimate the model $\hat{P}_{j,t} = \beta_0 + \beta_1 \hat{A}_{j,t} + \beta_2 \hat{Y}_{j,t} + \beta_3 \hat{U}_{j,t} + e_{j,t}$ using the `lm()` function.

```
# model_lm <- insert_code_here
# summary(model_lm)
```

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)

Overview. We want test if changes in the future price on the day of the WASDE report release are associated with changes in the WASDE forecast (e.g., acres, yield, or total use). In this model, P_t is the dependent variable and P_{t-1} is the main dependent variable. You will need to create three explanatory variables - $\Gamma^A_{j,t}, \Gamma^Y_{j,t}, \Gamma^U_{j,t}$ - that - remain at zero if there is no change in the forecast - take on a value of 1 if the forecast increased - take on a value of -1 if the forecast decreased

In your `wasde_wide` dataframe, use the `mutate()` function to create 3 indicator 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 $Acres_t - Acres_{t-1} > 0$)
- Variable = 0 if no change from previous month (e.g., if $Acres_t - Acres_{t-1} = 0$)
- Variable = -1 if change from previous month <0 (e.g., if $Acres_t - Acres_{t-1} < 0$)

```
# wasde_wide <- insert_code_here
```

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. Hint: You should have 2,980 observations in your `corn_wasde` dataframe. If you check your observations between September 9-14, 2010, you should see that only the September 10 row will have values coming from the `wasde_wide` data frame because that was the date that the WASDE report was released. Other days will contain NA. Your `corn_wasde` dataframe should contain 2980 observations.

```
# corn_wasde <- left_join(corn, wasde_wide, by = c("Date"))

# dim(corn_wasde)

# print observations between September 9-14, 2010. Sept 10 was the day the WASDE report was released.
# corn_wasde %>%
#   filter(Date >= "2010-09-07" & Date <= "2010-09-14")
```

Now you can run the event study model $P_t = \beta_0 + \beta_1 P_{t-1} + \beta_2 \Gamma^A_{j,t} + \beta_3 \Gamma^Y_{j,t} + \beta_4 \Gamma^U_{j,t}$ with the `lm()` function.

```
# eventstudy <- insert_code_here
# summary(eventstudy)
```

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?