problemset3

Brian Gilmore

2023-03-04

[1]

```
#Q1 Code
#Packages
library(pacman)
pacman::p_load(Ecdat, magrittr, dplyr, here, lubridate, ggplot2, fixest, lmtest)
# Reading data
ps3_df_raw = read.csv("C:/Users/gilmo/Downloads/ps3-data.csv")
# Some cleaning, and adding lagged values
ps3_df =
  ps3_df_raw |>
  select(date, unemployment, inflation, year, month, time) |>
  mutate(
    date = mdy(date),
    unemploy_lag = lag(unemployment),
    inf_lag = lag(inflation),
    yr_lag = lag(year),
    month_lag = lag(month),
    time_lag= lag(time))
nrow(ps3_df) == 241
```

```
## [1] TRUE
```

[2]

```
#Q2 Code
mod_s = lm(
  data = ps3_df,
  unemployment ~ inflation
)
summary(mod_s)
```

```
##
## Call:
## lm(formula = unemployment ~ inflation, data = ps3_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9414 -1.5142 -0.4412  1.3044  7.8128
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   7.8879     0.6689  11.792  < 2e-16 ***
## inflation    -0.9097     0.3165  -2.875  0.00441 ** 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 2.007 on 239 degrees of freedom
## Multiple R-squared:  0.03342,    Adjusted R-squared:  0.02938 
## F-statistic: 8.263 on 1 and 239 DF,  p-value: 0.00441
```

[3] Our estimated intercept expects that with 0 inflation, average unemployment is expected to be 7.8879, statistically different from 0 with a p-value < 2e-16. Our estimated coefficient on inflation tells us that we can expect a decrease of -0.9097 unemployment rate units with each unit increase of inflation rate. It is statistically significant at the 5% level that inflation is expected to decrease unemployment by -0.9097 for each unit increase in inflation with a p-value of .00441 < .05.

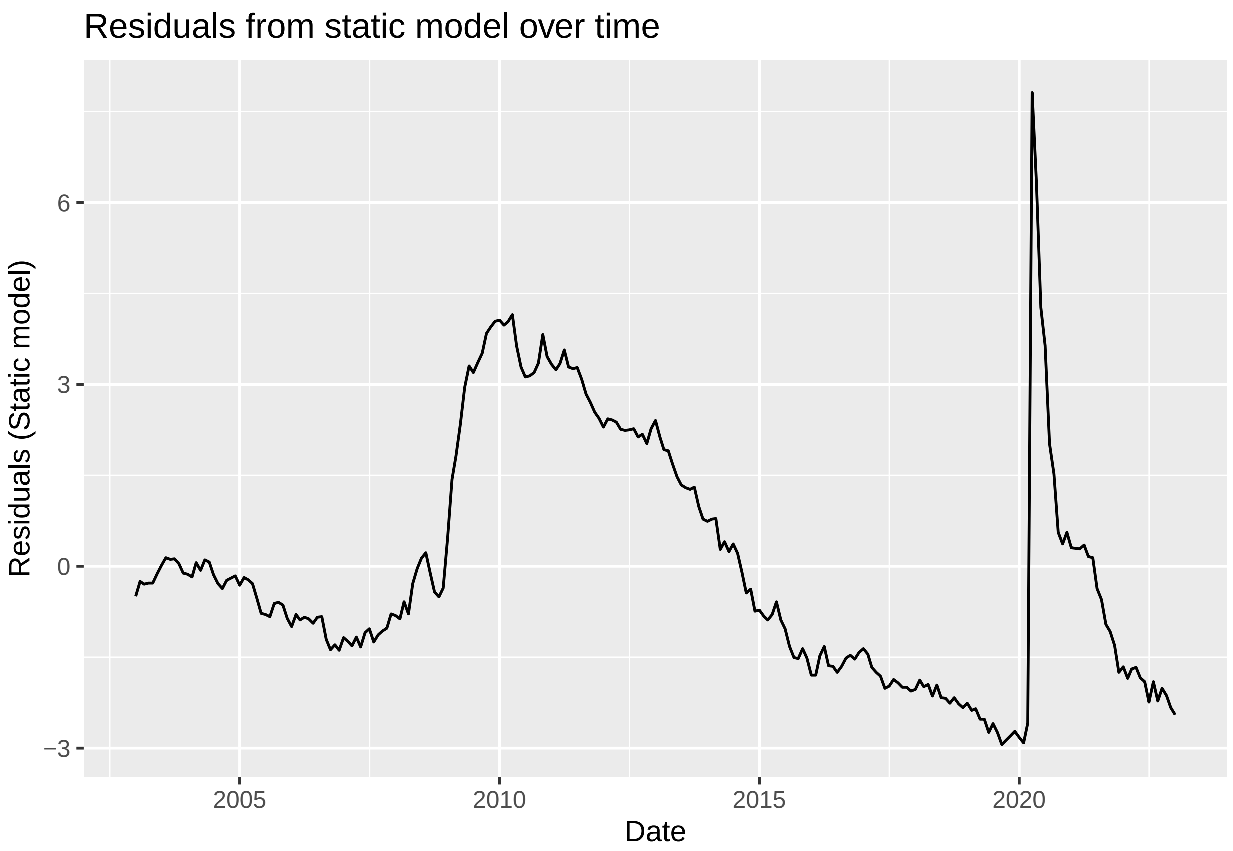
[4] In general OLS assumes no omitted variables, exogeneity, independent observations, homoskedasticity, normal distribution, and no multi-colinearity. More specfically, OLS unbiased coefficients depend on both assumptions of exogeneity. Our first assumption is that the disturbance is independent of the explanatory variables in the same period. The second assumption is that the disturbance is independent of the explanatory variables in the other periods. These assumptions are reasonable since we are looking at a static model that represents inflation’s effect on unemployment within the current period. Without these assumptions, our variable for inflation won’t have an immediate effect on unemployment and may have an effect on unemployment in the future. This would also indicate that current unemployment might depend on previous unemployment, since the disturbances are dependent on explanatory variables in other periods.

[5] With autocorrelation present in our static model, OLS will be inefficient and result in biased estimates for the standard errors. This could result in invalid statistical test results. However, our estimates for our coefficients will be unbiased. This is similar to heteroskedasticity’s effect on our model.

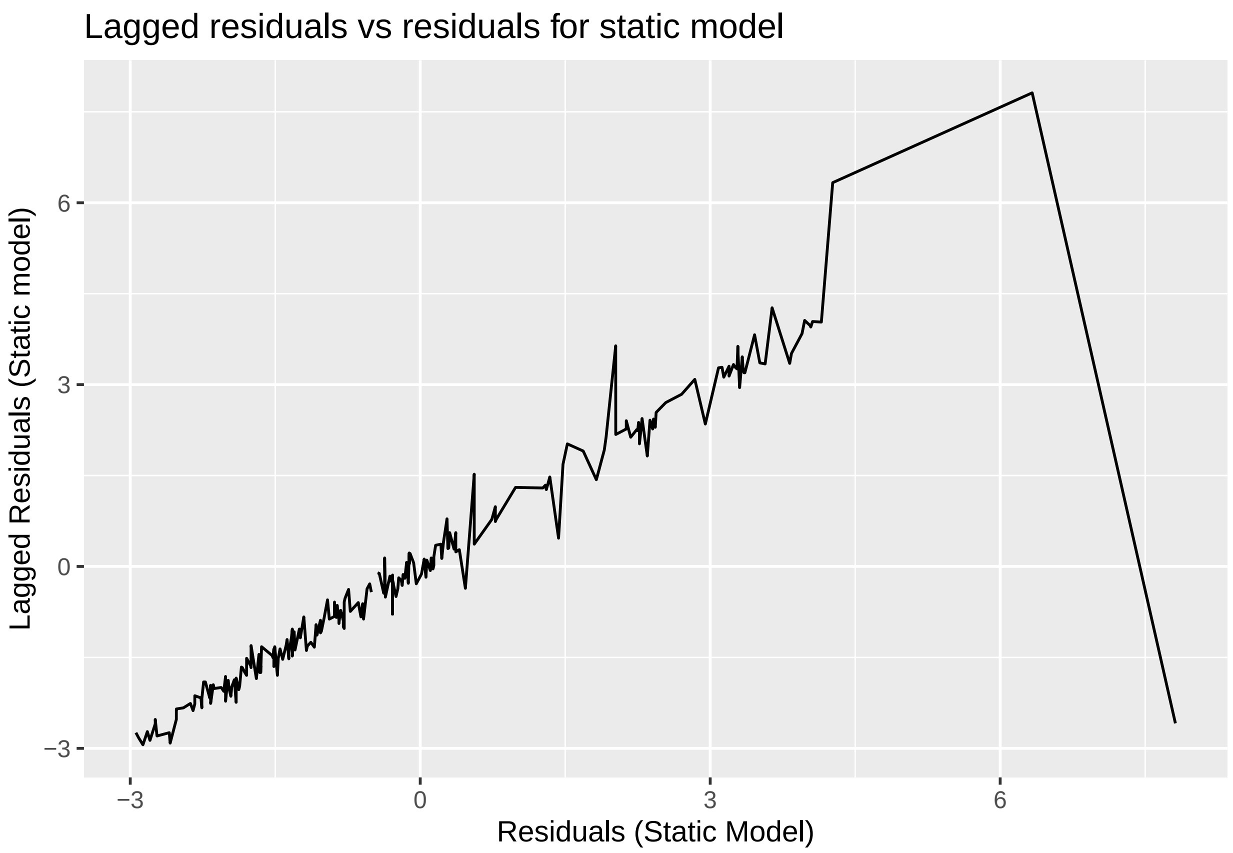
[6]

```r
#Q6 Code
#First we add our residuals and lagged residuals to our data
s_resid <- resid(mod_s)
s_resid_lag <- lag(s_resid)
ps3_df %<>% mutate(s_resid, s_resid_lag)

#Then we can plot s_resid to visualize autocorrelation
ggplot(data = ps3_df, aes(x = date, y = s_resid)) +
  geom_line() +
  labs(
    title = "Residuals from static model over time",
    y = "Residuals (Static model)",
    x = "Date")
```



```r
#Then we can plot s_resid_lag to visualize static model residuals over time
ggplot(data = ps3_df, aes(x = s_resid, y = s_resid_lag)) +
  geom_line() +
  labs(
    title = "Lagged residuals vs residuals for static model",
    y = "Lagged Residuals (Static model)",
    x = "Residuals (Static Model)")
```



```
#Looking at our plots, it seems there is some autocorrelation present.
#Lets check for AR(1) disturbance
mod_es = lm(
  data = ps3_df,
  s_resid ~ -1 + s_resid_lag)
summary(mod_es)
```

```
##
## Call:
## lm(formula = s_resid ~ -1 + s_resid_lag, data = ps3_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6570 -0.1929 -0.0522  0.1057 10.2339
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## s_resid_lag  0.93578    0.02335   40.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7222 on 239 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8705, Adjusted R-squared:  0.8699
## F-statistic:  1606 on 1 and 239 DF,  p-value: < 2.2e-16
```

From our test, there is statistically significant evidence that our coefficient on s_resid_lag is different from 0 at the 5% level with a p-value < 2e-16. This indicates that for our single lag, there is evidence of autocorrelation present, which means that we can reject the null hypothesis of no autocorrelation at the 5% level.

[7] In general, disturbances are autocorrelated when the shock from a disturbance in a certain period t correlates with nearby shocks in period t+1 and period t-1. In our model, there may be persistence present regarding our inflation variable. If there was a previous shock in inflation that had a large effect on unemployment, the lasting effects of inflation on unemployment might last for multiple periods where the disturbances will be correlated while unemployment is recovering. This is impossible to adjust for in a static model, since static models do not allow for persistence.

[8]

```r
#Q8 Code
lag_mod <- lm(unemployment ~ inflation + inf_lag, data = ps3_df)
summary(lag_mod)
```

```
##
## Call:
## lm(formula = unemployment ~ inflation + inf_lag, data = ps3_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0726 -1.5390 -0.4627  1.2556  7.3720
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2104     0.6706  12.243  < 2e-16 ***
## inflation     1.5790     0.9264   1.704  0.08961 .
## inf_lag      -2.6457     0.9255  -2.859  0.00463 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.981 on 237 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.06584,    Adjusted R-squared:  0.05796
## F-statistic: 8.352 on 2 and 237 DF,  p-value: 0.0003125
```

[9] In our regression the intercept coefficient represents that we expect average unemployment to be 8.2104 when inflation and inf_lag are 0. This estimate for the intercept is statistically significant at the 5% level. Our estimate for the coefficient on inflation expects that for each unit-increase in the current period t, we can expect average unemployment to increase by 1.5790. This estimate is not statistically different from 0 at the 5% level as the p-value > .05 at 0.08961. The estimated coefficient for inf_lag is -2.6457 which tells us that we can expect a decrease of 2.6457 in average unemployment for each unit-increase in the inflation rate for the previous period, t-1. This estimate is statistically significant at the 5% level with a p-value at 0.00463.

[10] With a lagged explanatory variable, OLS is inefficient and has biased standard error estimates. However, our coefficient estimates are still unbiased, unless we were to add a lag of our outcome variable.

[11] In this context, it makes more sense to use a dynamic model since we want to evaluate the effect of inflation on average unemployment in the previous period, as well as in the current period. Dynamic models allow for persistence, which helps us understand the behavior of variables in previous periods. The previous period of inflation may explain our estimates better since inflation rates could have persistent effects. With

a static model observing only the current period, it would be harder to determine the effects of inflation on unemployment without looking at inflation in previous periods. If we were to add a lag of unemployment at t, unemployment at t-1 and our disturbance at period t would both depend on lagged disturbance at period t-1. This means that our regressor is correlated with its contemporaneous disturbance. With this correlation, autocorrelation is present as our model violates contemporaneous exogeneity. This would also create bias and inconsistency in our coefficients.

[12]

```
mod_d <- lm(unemployment ~ inflation + inf_lag + unemploy_lag, data = ps3_df)
summary(mod_d)
```

```
##
## Call:
## lm(formula = unemployment ~ inflation + inf_lag + unemploy_lag,
##     data = ps3_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9777 -0.1809 -0.0253  0.1455  9.8497
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.10781    0.29723   3.727 0.000242 ***
## inflation    -0.14098    0.33339  -0.423 0.672787
## inf_lag      -0.19256    0.33589  -0.573 0.567000
## unemploy_lag  0.92913    0.02305  40.304  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7071 on 236 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8815, Adjusted R-squared:   0.88
## F-statistic: 585.2 on 3 and 236 DF,  p-value: < 2.2e-16
```

[13] The estimates are different in this model since all of our coefficients are biased.We cannot separate the effects of period t disturbance on unemployment at t from unemployment at t-1 on unemployment at t. Both disturbance at t and unemployment at t-1 depend on disturbance at t-1, therefore every other coefficient is biased. We also see that our coefficient estimates for inflation and inf_lag are not statistically significant.

[14] OLS is biased and inconsistent for the coefficient estimates and we are violating contemporaneous exogeneity where Cov(x_t, u_t) doesn’t equal zero. This is due to autocorrelated disturbances present in our regression with a lagged outcome variable.

[15]

```
#Q15 Code
d_resid <- c(NA, resid(mod_d))
bgreg <- lm(d_resid ~ inflation + inf_lag + unemploy_lag + lag(d_resid) + lag(d_resid, 2), data = ps3_d
waldtest(bgreg, c("lag(d_resid)", "lag(d_resid, 2)"))
```

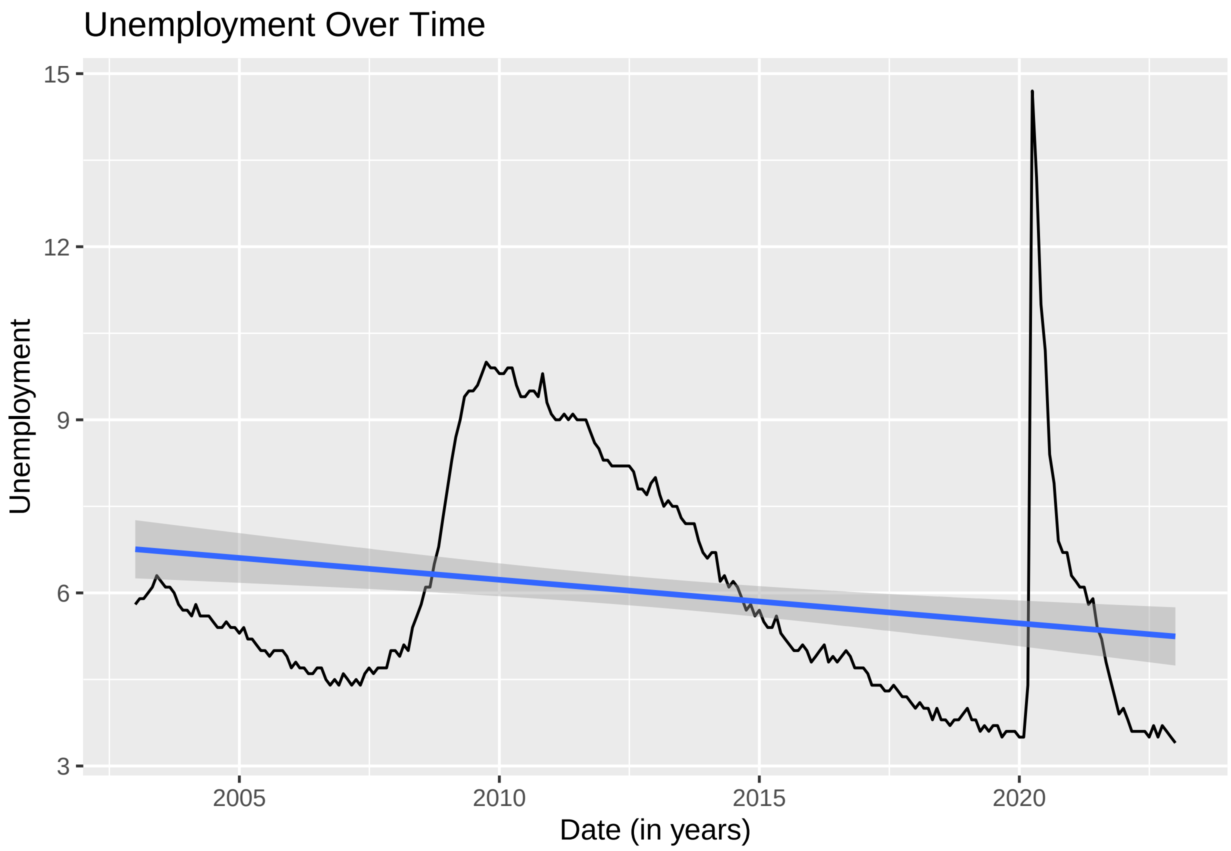
```
## Wald test
##
## Model 1: d_resid ~ inflation + inf_lag + unemploy_lag + lag(d_resid) +
##      lag(d_resid, 2)
## Model 2: d_resid ~ inflation + inf_lag + unemploy_lag
##   Res.Df Df      F Pr(>F)
## 1    232
## 2    234 -2 1.1526 0.3176
```

We fail to reject the null hypothesis that no autocorrelation is present since our p-value is .3176 $<$ .05.

[16]

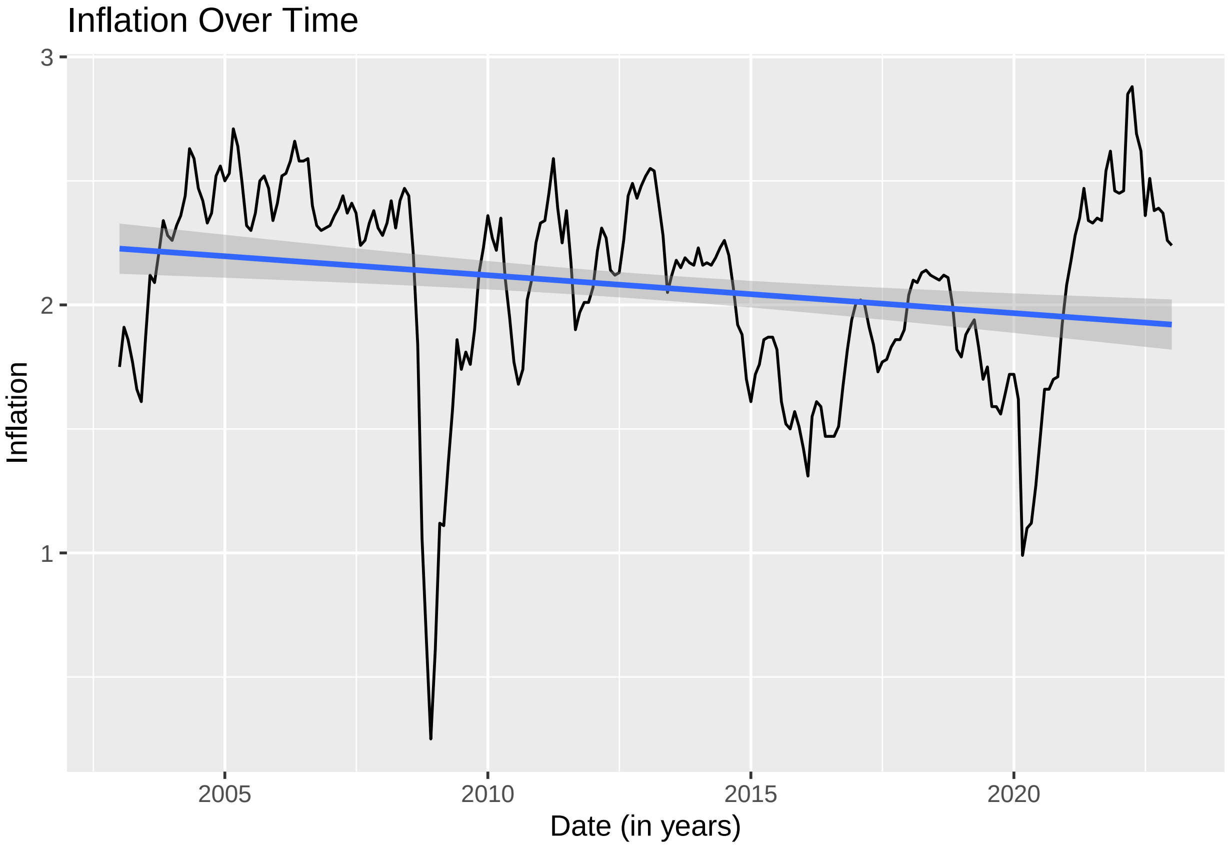
```
#Q16 Code
#First the plot of Unemployment vs Time:
ggplot(data = ps3_df, aes(x = date, y = unemployment)) + geom_line() +
    labs(title = "Unemployment Over Time",
    y = "Unemployment",
    x = "Date (in years)") + geom_smooth(method = "lm")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
#First the plot of Unemployment vs Time:
ggplot(data = ps3_df, aes(x = date, y = inflation)) + geom_line() +
    labs(title = "Inflation Over Time",
    y = "Inflation",
    x = "Date (in years)") + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



[17] No, both variables appear to be non-stationary. We can see that throughout both plots there are large fluctuations in the data over time. In both distributions the mean looks to be dependent on time as our regression lines in both are downward sloping. We also see that the variance is dependent on time as there are times where both variables fluctuate wildly at different time periods. It also appears that in both distributions that the covariance between instances of variables is not constant, where both instances of variables at different times depend on time, not the difference between time and an arbitrary constant. All of these requirements are violated in our variables.

[18] Non-stationary variables could lead to fake/not valid spurious results due to the violation of stationary variance. This would lead to regressions that are falsely statistically significant. This also implies that we cannot trust OLS as it would result in false rejection of the null hypothesis in running our analysis. In our regressions above, estimates that we might have found to be statistically significant may be false and give us incorrect results for our models due to the violation of stationary requirements.

[19] If we find that random walks are present through the Dickey-Fuller tests, we may be able to combat non-stationary variables by implementing difference stationary processes. This involves adjusting disturbances to account for non-stationarity. For example we could take our disturbance at a specific period t, and subtract the disturbance at period t-1 to get better representation of the change in our disturbances over time. This would potentially bring our model back to a stationary requirement if random walks are present.

[20] Yes we should be concerned about reverse causality in this setting since there is a possibility for there to be a causal relationship between inflation and unemployment. If there is a correlation found between the two variables and our assumption of exogeneity is present, unemployment could cause inflation. If we were to regress unemployment on the lag of inflation, our results may prove to be statistically significant, but the results may violate exogeneity and prove to be biased or produce spurious results. By only regressing with a lagged variable, there is room for omitted variable bias, which wouldn't tell us much about the causal effect

of unemployment on inflation. Ideally to accurately measure the causalty, we would be able to set up an experiment.