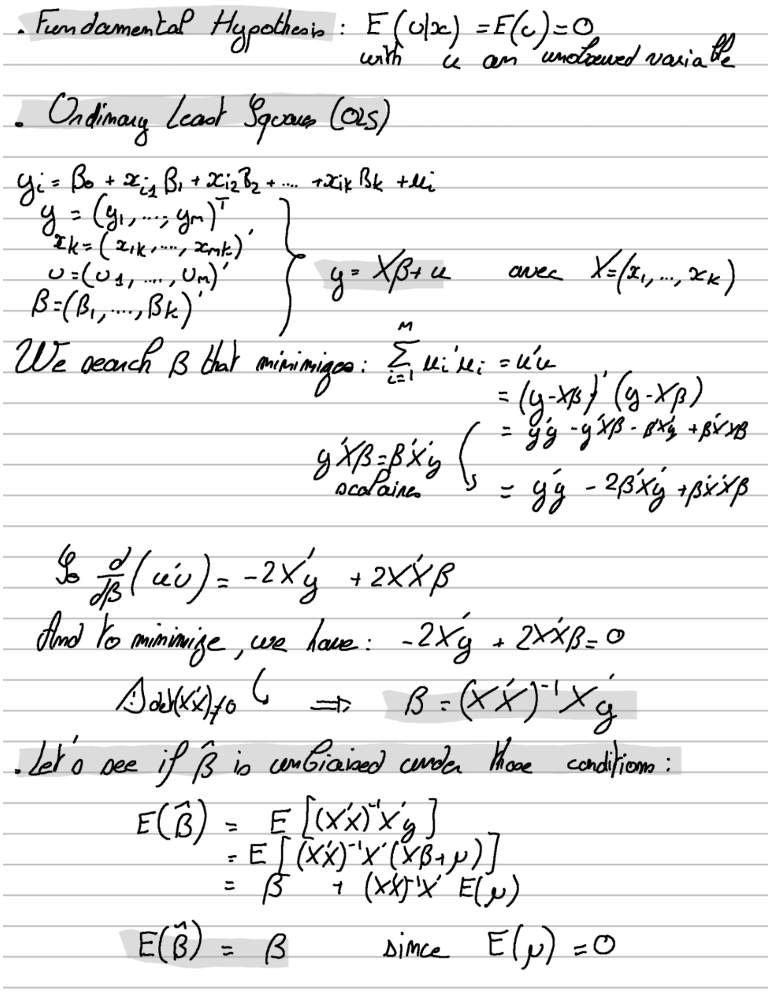
ECONOMETRICS PROJECT

PART 1 : CROSS-SECTION DATA

1. Fundamental hypothesis :

   Unobserved variables have zero means, the conditional mean of the error is equal to the unconditional mean, X and u are uncorrelated so E(u)=E(u|X)=0.

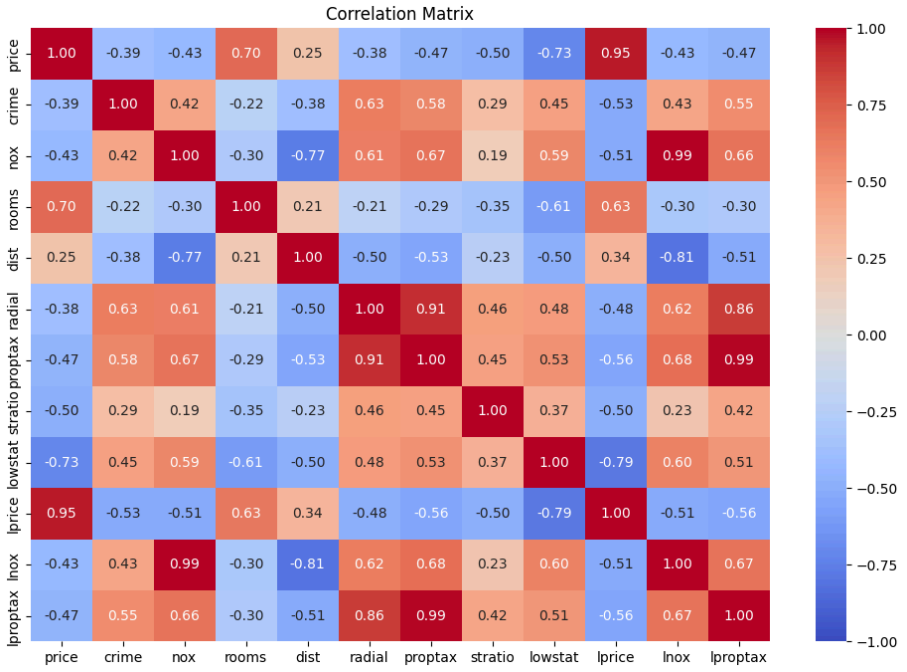
2.



3. During WW2, British engineers chose to reinforce their bomber planes based on where they were shooted on the most (they looked at the bullet holes on their cabins). However, Abraham Wald remarked that there was a survival bias, they were only looking at the cabins of planes that hadn't been shooted down. Consequently, they needed to reinforce the parts of the planes that weren't armed on the remaining planes.

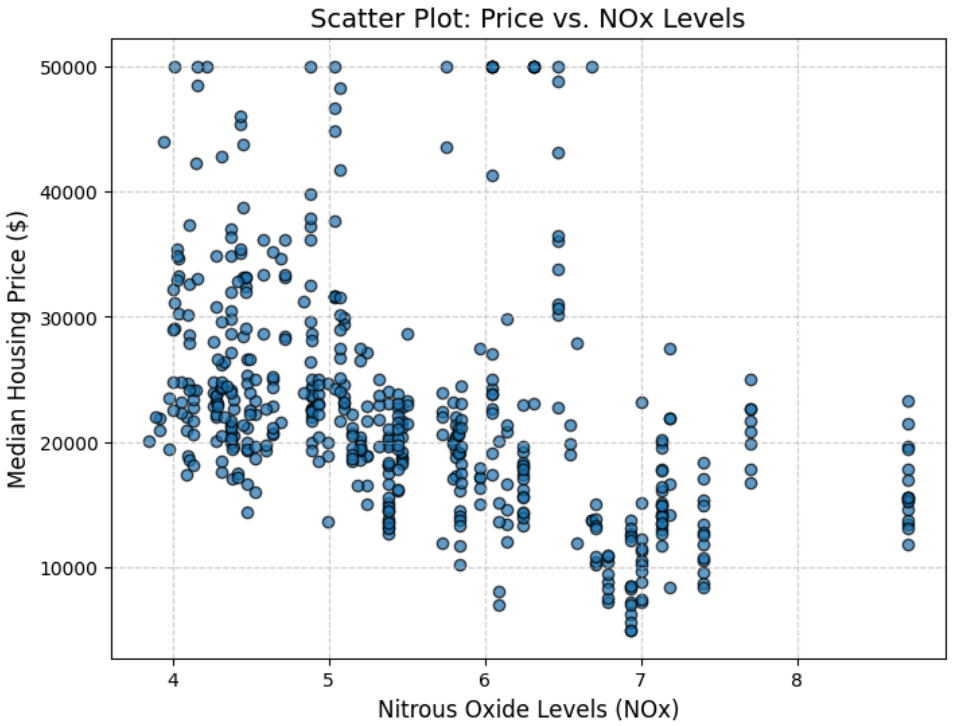
4. Sample selection bias occurs when unobserved variables, like soil quality in the example from the course, affect the outcome being measured. In studying the effect of fertilizers on crop yields, factors like climate, soil quality, and pests also influence yields. If we select plots systematically, such as applying more fertilizer to high-quality soil, the relationship between fertilizer and yield becomes confounded, as soil quality is unobserved. Randomly selecting plots helps avoid this bias by ensuring the sample better represents the population, isolating the true effect of fertilizer.

5. Multicollinearity occurs when two or more columns of X'X are nearly linearly dependent, causing det(X′X) to approach zero, making it difficult to calculate the inverse of X′X. This hampers the computation of standard deviations and affects model reliability (cf question 2). Solutions include adding more observations or removing highly correlated variables. Small changes in the data can have a big impact on the estimator.
   Lproptax and Lnox are collinear with proptax and nox, which is normal, however radial has a high VIF and is nearly collinear with proptax, I think we can neglect this but we should be cautious with this variable. We don't have any other issue of multicollinearity.



6. Cf .ipynb file.

7. mean_low : 27 000 $
   mean_medium : 19 800 $
   mean_high : 17 900 $
   We can conclude that, on average, the less an area is polluted , the higher is the median housing price of the area.

8. 

   This is not a Ceteris Paribus effect since the other variables such as crime, etc aren't fixed.

9. **constant : -18682.26**
   The constant doesn't have a real world interpretation, however it is amusing to note that it is a negative value. It just represents the value of price when all other variables are set to 0.

   crime : -136.54
   *For every **1-unit increase in the crime rate (crimes per capita)**, the housing price decreases by approximately **$136.54**, holding all other variables constant. Higher crime rates have a negative impact on housing prices, as expected.*

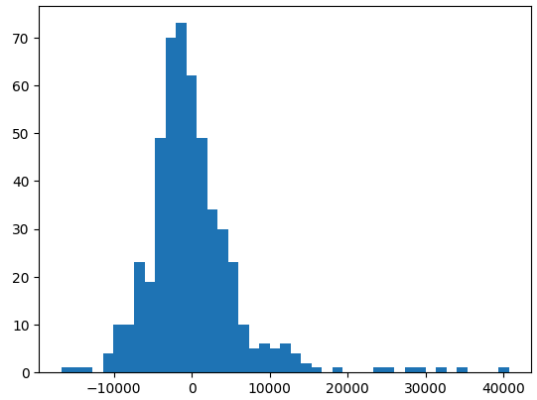
   nox : -660.4672465241154
   *For every **1-unit increase in nitrous oxide levels**, the housing price decreases by approximately **$660.47**, holding all other variables constant. Areas with higher pollution levels (indicated by nox) tend to have lower housing prices. This reflects*

that people generally prefer cleaner environments and are willing to pay less for homes in polluted areas.

rooms : 7797.928603057132
*For every **additional room** in a house (on average), the housing price increases by approximately **$7797.93**, holding all other variables constant. Larger homes with more rooms are associated with higher housing prices, reflecting that size is a key determinant of property value.*

proptax : -89.41439612639125
For every **1000$ increase in property tax**, the housing price decreases by approximately **$89.41**, holding all other variables constant.Interpretation: Higher property taxes are associated with lower housing prices, possibly because buyers factor in the additional ongoing cost of owning the property when deciding how much they are willing to pay.

It is important to note that those factors represent a correlation and not necessarily a causality.



We can see that there is no bias in the model. The histogram represents a gaussian centered on 0. However the variance seems to be quite big. Most of the errors are between -10k $ and 10k $.

10. **Model :**
$\ln(price) = 8.655 - 0.0125 \cdot crime - 0.0476 \cdot nox + 0.2816 \cdot rooms - 0.0043 \cdot proptax$

Constant (8.655):

> *The intercept represents the expected value of ln(price) when all independent variables are zero.*

Crime (−0.0125):

A one-unit increase in the crime rate decreases the log of the price by 0.0125 holding other factors constant. In percentage terms, this corresponds to approximately a 1.25% decrease in housing prices for each crime committed per capita.

Nox (−0.0476):

A one-unit increase in NOx emissions reduces the log of the price by 0.0476, equivalent to a 4.76% decrease in housing prices, holding other variables constant. This reflects the negative impact of pollution on housing values.

Rooms (0.2816):

An additional room increases the log of the price by 0.2816, translating to approximately a 28.16% increase in housing prices, all else being equal. This indicates that larger homes are significantly more valuable.

Proptax (−0.0043):

A one-unit increase in property tax rate reduces the log of the price by 0.0043, or a 0.43% decrease in housing prices, ceteris paribus. Higher taxes negatively affect property values.

11. **Model :**
ln(price)=9.7505−0.0128 · crime−0.2769 · ln(nox)+0.2802 · rooms−0.1779 · ln(proptax) Constant (9.7505):

The intercept represents the expected value of ln(price) when all independent variables (crime,etc) are zero. This is a scaling factor.

Crime (−0.0128):

It's the same as the precedent model.

ln(nox) : (−0.2769):

A 1% increase in NOx emissions (ln(nox)) is associated with a 0.2769% decrease in housing prices, holding other variables constant. This reflects a significant negative impact of pollution on property values.

Rooms (0.2802):

It's the same as the precedent model.

ln(proptax) (−0.1779):

A 1% increase in property tax (ln(proptax)) reduces the price by 0.1779%. This indicates that higher taxes reduce property values.

12. Coefficient for nox: -660.4672465241154
    Standard error for nox: 314.63381312692934
    p-value for nox: 0.036302434683150066

    The p-value is greater than 0,01 so **we fail to reject the H0 hypothesis** which is Bnox = 0.

13. We rewrite the model according to theta = Bcrime - Bproptax, we therefore have a new X (see the notebook).
    We obtain a p-value that is equal to = 0.
    We therefore reject hypothesis H0 : Bcrime is different from Bproptax.

14. To answer this question, we've chosen to use the F-value method with a restricted model and an unrestricted one.
    We've found :
    - F-statistic: 19.311290349552785
    - Critical F-value at 10% significance level: 2.3132002236303255
    - F-statistic > Critical F-value
    Consequently we should reject the hypothesis at the 10% level : Bnox and Bproptax aren't both equal to 0.

15. We obtain a p-value of 0.86.
    We can't reject the H0 hypothesis. Consequently we can't deny that b_{nox} = -500 and $b_{proptax}$ = -100 at the 10% level.

16. To answer this question, we've chosen to perform a Chow test. We have the following results :
    - F-statistic: 5,57
    - p-value: ~0
    Consequently, we can reject the hypothesis that all coefficients are the same for observations with low levels of *nox* vs. medium and high levels of *nox*.

17. cf the notebook to see the code.
    The H0 hypothesis is : - βnox,low = βnox,high and
    - βproptax,low = βproptax,high
    - All the other coefficients are equals
    We obtained a p-value of 0,458, consequently, we fail to reject the H0 hypothesis at

the 10% level.

PART 2 : HETEROSCEDASTICITY

18. We have heteroskedasticity when the variance of the errors is not constant and changes depending on the values of the explanatory variables. It is a problem because the estimators remain unbiased, but they're not efficient anymore. Especially for OLS estimators, they don't give the smallest variance anymore among the BLUE: Best Linear Unbiased Estimators.
    Another problem is about standard errors. With heteroskedasticity, those errors can be over or underestimated, leading to unreliability on confidence intervals.

19. F > F_critique. We reject the hypothesis of homoscedasticity.

20. F > F_critique. We reject the hypothesis of homoscedasticity.

21. F > F_critique.We reject the hypothesis of homoscedasticity.

22. We notice that F is the same for question 19 and 21. It can be linked to the fact that, in q.19, we choose to go over price with nox and proptax, and in q.21, we're using log(nox) and log(proptax) to go over log(price).

23. Proptax is the most significant variable causing heteroskedasticity. Using proptax as weight and running a WLS regression, we get smaller standard errors than the errors we had with the simple OLS regression.

PART 3 : TIME SERIES DATA

/!\ For this part we used the CPI value as the inflation rate since we didn't find

any inflation column in the given file.

24. We have strict stationarity when the distribution of the series does not change over time. This means that the joint probability distribution of any set of time points remains the same when we add a delay over time.
    On the other side, we have weak stationarity if we have a constant mean, a constant variance and the covariance only depends on the delay : $Cov(Xt,Xt+h)=\gamma(h)$.

25. Ergodicity means that the process forgets its initial conditions: the autocorrelation of order k tends to 0 when k gets close to infinity.
    The ergodic theorem states that, if y_t is strictly stationary and ergodic, and its mean is finite when t gets close to infinity, then the time average converges to the spatial average (the mean).

26. Stationarity is needed in this theorem because it ensures that the mean and the variance do not change over time.
    Ergodicity is needed because it ensures that the process goes over its entire state space over time. Without ergodicity, a stationary process could have a strong dependence across time and this would prevent the time average from converging to a finite value.

27. Spurious regression happens when there is a statistical relationship in regression analysis, where two or more variables are related even though they are not causally connected. This can lead to incorrect conclusions about relationships between variables and need to be avoided.

28. cf the notebook.

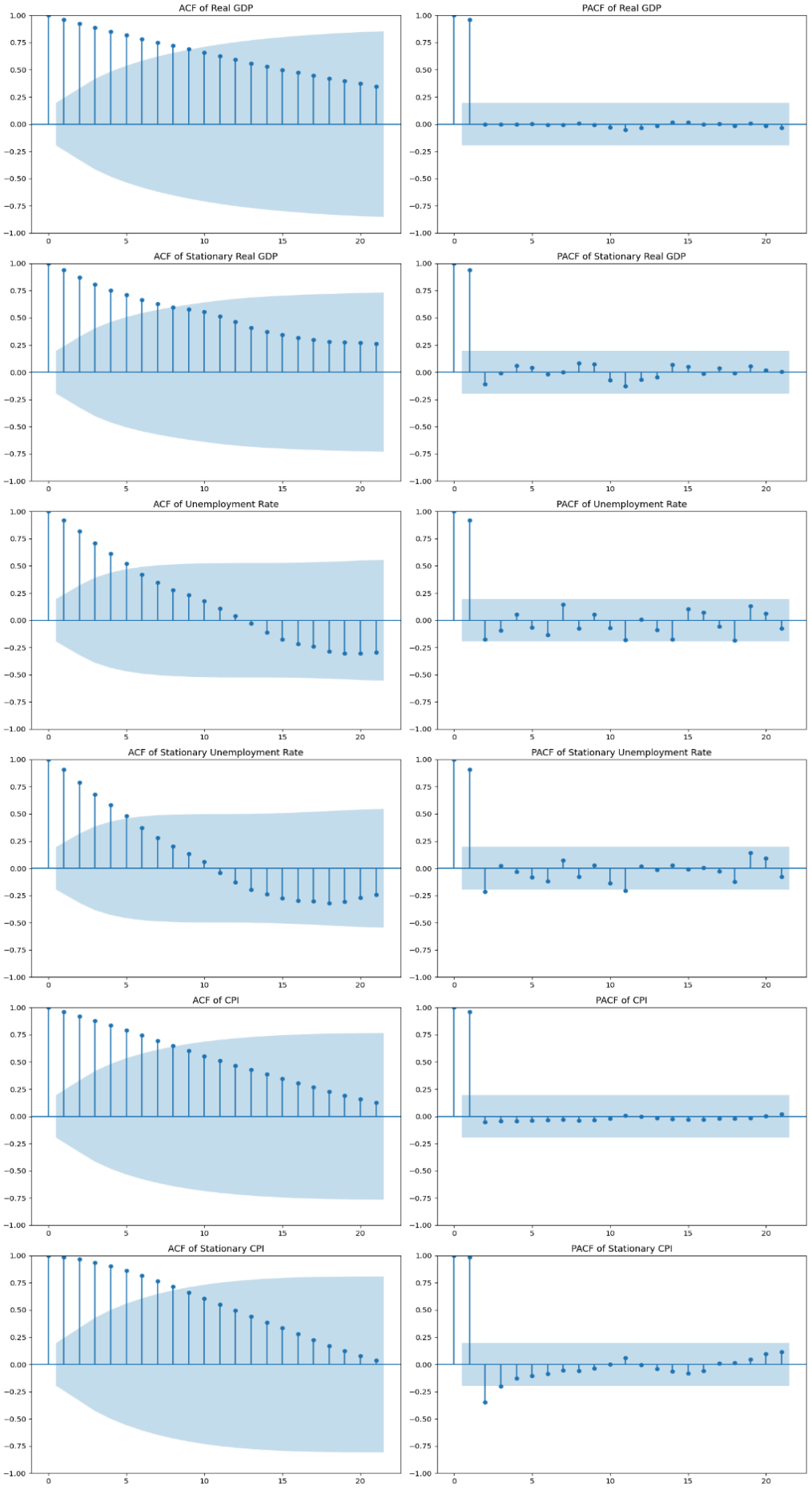
29. We obtain :
    - for the real GDP : p-value = 1.0
    - for the Unemployment Rate : p-value = 0.15
    - for the CPI : p-value = 0.49

    Consequently, we aren't able to reject the H0 hypothesis at the 5% or 10% levels, the series are non-stationeries.

30. Since the series are non-stationeries, which seems logical (GDP and inflation growth) we will use the series from the 28th question.

31. The ACF measures the correlation between X_t and X_{t-k}, where k is the delay, considering all intermediate delays.
    The PACF measures the correlation between X_t and X_{t-k}, adjusted for the correlations with all other delays (1,2,…,k−1).

32. For the non-stationary variables, we observe that the ACF has a slow decrease and that there's a peak for the first lag of the PACF, then it is close to zero.
For the stationary variables, we note that the ACF has a decrease less linear that can oscillate and, on the PACF, we observe that there's still a peak at the beginning and the other values are very low.


ACF of Real GDP
PACF of Real GDP
ACF of Stationary Real GDP
PACF of Stationary Real GDP
ACF of Unemployment Rate
PACF of Unemployment Rate
ACF of Stationary Unemployment Rate
PACF of Stationary Unemployment Rate
ACF of CPI
PACF of CPI
ACF of Stationary CPI
PACF of Stationary CPI

33. The principle of parsimony says that, when we have several models for a certain set of data, the simplest one should be chosen (among the ones that describe the data well). Ockham's razor, a philosophical idea attributed to William of Ockham, states that "Entities should not be multiplied beyond necessity."
    In terms of modeling, this means avoiding unnecessary complexity by choosing models that are simple yet effective in explaining or predicting outcomes.
    This is formalized by information criteria. They help find a compromise between a good description of the data and a reasonable complexity.

34. Autocorrelation of errors happens when the residuals errors from a regression model are not independent of each other but are correlated across observations. This violates an assumption of the classical linear regression model (assumption that the error terms are independently distributed). This leads to biased standard errors and thus, the estimators are unreliable.

35. We compute the Durbin Watson statistic. We get $0.131 < 2$. So, there is a positive autocorrelation.

36. For this question, we conducted additional research online and identified the Cochrane-Orcutt transformation as a potential solution. We applied this method to two types of autocorrelation structures : AR(1) and AR(2). The AR(1) model proved more suitable for our dataset.
    Initially, the Durbin-Watson statistic was around 0.13. After applying the Cochrane-Orcutt transformation, it rose to approximately 1.124. Since values closer to 2 indicate less autocorrelation, we successfully reduced the problem, though we did not fully eliminate it.
    We think it may be possible to fully eliminate autocorrelation with other tools than GLS.

37. Cf the notebook.

38. The number of observations is reduced by 2. Because of lag 2, we cannot use all the data, we lose 2 observations.

39. The no-Granger causality hypothesis tests whether lagged values of unemployment provide statistically significant information about the future values of GDP.
    We get that the unemployment rate does not Granger-cause GDP, meaning that the lagged values of the unemployment rate do not provide significant predictive power for future values of GDP.

40. We see with the graph that coefficients are not stable. We verify that intuition with the Chow Test that gives us a p-value of $1.1102230246251565e\text{-}16 < 0.05$. We reject the hypothesis of stability.

41. We get that the p-value is 3.2834157843698386e-06 < 0.1. So it means that there is evidence of instability in the coefficients of the regression model over time.

42. We test the coefficients between the 2 periods (1933-1965) and (1966-1998). After the Chow Test, we get the p-value of 1.1102230246251565e-16 < 0.05.
We reject the hypothesis, the coefficients are not equal.