Time Series Analysis Spring 2021

Midterm Exam – March 2, 2021

Instructions

i. Answer all 10 questions.
ii. You may type your work or write it all by hand, or a combination. If you write it all by hand, make sure to write neatly. I prefer you type your explanations, but if you type a lot slower than you can write neatly, don't worry about it. If you decide to type most of it, feel free to write the equations by hand—only type them if you are a master and can type them as fast as you can write them neatly.
iii. Show and explain your work where relevant. A correct final answer, without explanation, is worth little. A good explanation and mostly correct work, with a wrong final answer due to a small mistake, is worth a lot.
iv. Be neat and concise.
v. List and explain any assumptions you make.
vi. You may use any resources available other than speaking or otherwise communicating with anyone else.
vii. Keep your camera, microphone, and speakers on for proctoring and for test related communications.
viii. Upload an electronic copy of your work to canvas by 7:00 PM.

Background

You are interested in determinants of housing prices in Florida. You have the following three hypotheses:

i) Building permits mean an anticipated increase in future supply. So when building permits increase, you expect a decrease in price shortly thereafter.
ii) The cost of borrowing to build or buy a home depends on the real interest rate, Interest. This is the difference between the interest rate and expected inflation. So, when interest rates increase, you expect housing prices to fall shortly thereafter.
iii) When inflation is expected to be higher, home prices go up accordingly. So, when inflation expectations increase, you expect an increase in home prices shortly thereafter.

You have monthly data on the median per square foot price of homes listed for sale in Florida since July of 2016. You plan to ignore data after December of 2019 for this purpose. Assume, for purposes of this question, that gathering more data from earlier periods is not possible. (In reality, it exists, it is just not readily available for free and on FRED.) You will need to keep in mind the limitations imposed on model complexity by the amount of data. In particular, you do not have enough data to estimate models with many lags for multiple variables while keeping a reasonable number of degrees of freedom.

Given the limited data, you have in mind the following causal model:

$$\ln List_t = \beta_0 + \beta_P \ln Permits_{t-1} + \beta_{Int} \ln Interest_{t-1} + \beta_{Inf} \ln Inflation_{t-1} + r_t .$$

A Stata log file, augmented with several graphs, accompanies this exam.

Midterm Exam – March 2, 2021

Questions

1) Express the model above in first differences. Under what conditions would you need to work with the differenced model instead of the original?

If lnList is I(1), or simply highly persistent with a first order autocorrelation parameter near 1, first differencing is necessary to avoid spurious correlations.

$$\Delta \ln List_t = \beta_P \mathrm{H} \ln Permits_{t-1} + \beta_{Int} \Delta \ln Interest_{t-1} + \beta_{Inf} \Delta \ln Inflation_{t-1} + \Delta r_t$$

2) Suppose you think the residual, r_t, follows an AR(1) process with parameter ρ. Write the dynamically complete version of the original model ***and*** of the model in first differences.

Without differencing:

$$\begin{aligned} r_t &= \rho r_{t-1} + \varepsilon_t \\ r_{t-1} &= \ln List_{t-1} - \beta_0 - \beta_P \ln Permits_{t-2} - \beta_{Int} \ln Interest_{t-2} - \beta_{Inf} \ln Inflation_{t-2} \\ r_t &= \rho \ln List_{t-1} - \rho\beta_0 - \beta_P \ln Permits_{t-2} - \rho\beta_{Int} \ln Interest_{t-2} - \rho\beta_{Inf} \ln Inflation_{t-2} + \varepsilon_t \\ \ln List_{t-1} &= \beta_0 (1-\rho) + \beta_P \ln Permits_{t-1} + \beta_{Int} \ln Interest_{t-1} + \beta_{Inf} \ln Inflation_{t-1} \\ &\quad + \rho \ln List_{t-1} - \rho\beta_P \ln Permits_{t-2} - \rho\beta_{Int} \ln Interest_{t-2} - \rho\beta_{Inf} \ln Inflation_{t-2} + \varepsilon_t \end{aligned}$$

With differencing:

$$\begin{aligned} \Delta \ln List_{t-1} &= \beta_0 (1-\rho) + \beta_P \Delta \ln Permits_{t-1} + \beta_{Int} \Delta \ln Interest_{t-1} + \beta_{Inf} \Delta \ln Inflation_{t-1} \\ &\quad + \rho \Delta \ln List_{t-1} - \rho\beta_P \Delta \ln Permits_{t-2} - \rho\beta_{Int} \Delta \ln Interest_{t-2} - \rho\beta_{Inf} \Delta \ln Inflation_{t-2} + \Delta\varepsilon_t \end{aligned}$$

The remaining questions refer to the Stata do file and (augmented) log file provided. The file has page numbers, and line numbers that start at 1 on each page.

3) What is the purpose of the commands on page 1 lines 25-32?

These lines make sure Stata properly recognizes time in the data. A monthly date is set for time related calculations such as lags and differences and a categorical variable is created for month of year (1-12) for capturing seasonal effects.

4) What is the purpose of the commands and results from page 2 line 4 through page 3 line 30, and what conclusion should be drawn from the results of these commands?

The purpose is to determine whether lnList is exhibits high persistence or only weak dependence, since further analysis required the time series be stationary and weakly dependent. The AC and PAC are consistent with an I(1) process and indicate very high persistence, and the Dickey Fuller test cannot reject the null hypothesis of an I(1) process. The conclusion is that this series demonstrates high persistence and should be differenced before further analysis.

Midterm Exam – March 2, 2021

5) Four sets of models are estimated. What are the differences between the sets (***not*** between the models in a given set)? Which set is better for the purpose at hand? Why?

The differences are a) whether month dummies and a time trend are included and b) whether the data is differenced before estimating the model. The PAC and AC and Dickey Fuller test discussed previously indicate differencing is necessary. Seasonal indicator variables should be used to deal with purely seasonal effects (e.g. weather impacts on home building) unless there is a clear reason not to. Hence, set 4, which does both, is the best of these.

6) There are three models within the set you chose. Each of those is estimated twice. What is the difference between the two sets of estimates? Does the difference matter? Why? Which is better? Why?

For each model in the set, the first estimate uses the command ***regress*** which calculated default standard errors that assume no serial correlation and no heteroskedasticity in the residuals, while the second uses the command ***newey*** which calculates standard errors robust to autocorrelation and heteroskedasticity in the residuals. Thus, unless you are very confident your model has no autocorrelation or heteroskedasticity in the residuals, use the second estimate with Newey-West standard errors. Note the Breusch-Godfrey tests suggest autocorrelation remains.

7) For the set you chose as best, interpret the F-test for the first model in that set. That is, if set X is best, interpret the F-test that follows one of the two estimates of Model X.1. Again, there are two versions. Use the better one. Your answer to 6 should have made it clear what the difference is, which is better, and why.

This is a test (using the ***testparm*** command for testing sets of parameters) of the null hypothesis that neither the first lag of inflation nor the first lag of the interest rate have predictive power. The second version is best because it uses calculations robust to autocorrelation and heteroskedasticity. The p-value of 0.3545 indicates there is very little evidence upon which to conclude these variables are predictive of list prices.

8) How do the three models in your chosen set relate to the model set out on the previous page and to questions 1 and 2?

All are derived from the model set out in the background section. Model 4.1 is in first differences, like in question 1. Model 4.2 is the second equation in question 2, which would be dynamically complete if the residuals of the baseline model were a simple AR(1) process. The third model simply deleted inflation and interest as predictors from Model 4.2 following the test that shows no evidence they contribute predictive power. (If you incorrectly chose a non-differenced set, this would differ slightly)

9) What assumption must be defended to apply a causal interpretation to the results of this model, as opposed to a purely predictive one?

No omitted causes of list price are contemporaneously correlated with permits, interest, or inflation.

10) Within the set of models you chose as best, X, which model is best for predicting *List*? That is, X.1, X.2, or X.3? Why? Which is best for testing they hypotheses of interest? Why? If the two are different, why?

For prediction, one could make an argument for either model 4.2 or model 4.3. Model 4.3 is more parsimonious, dropping variables that seem not to have predictive power. But if the content knowledge indicating they should be controlled for is strong, Model 4.2 is better and 4.3 is simply overfit to the data. Model 4.1 is not as useful because it lacks lagged variables that we see have predictive power.

For causation, the three null hypotheses are that the three coefficients in the original model are zero. The alternative hypotheses are that the coefficient on permits is negative, the coefficient on interest rates is negative, and that the coefficient on inflation is positive. Model 4.3 does not contain all three coefficients, so it simply cannot test these hypotheses. As long as we use the Newey-West standard errors, and can defend the assumption discussed in question 9, we can make an argument for either model 4.1 or model 4.2. The coefficients of the first model are a bit more precisely estimated because the second lags of the three predictors in Model 4.2 don't appear to add anything useful, so I would probably take Model 4.1, which is the direct application of the hypothesized model after differencing. But, if you chose 4.2, that is not a bad answer.