Gus Lipkin - CAP 4763 Time Series Midterm

> **Corrections are marked by underlines**

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

- $List_t = \beta_0 + \beta_P lnPermits_{t-1} + \beta_{Int} lnInterest_{t-1} + \beta_{Inf} lnInflation_{t-1} + r_t$
- $\Delta List_t = \beta_0 + \beta_P \Delta lnPermits_{t-1} + \beta_{Int} \Delta lnInterest_{t-1} + \beta_{Inf} \Delta lnInflation_{t-1} + \Delta r_t$
- $\underline{\Delta List_t = \beta_0 + \beta_P \Delta lnPermits_{t-1} + \beta_{Int} \Delta lnInterest_{t-1} + \beta_{Inf} \Delta lnInflation_{t-1} + \Delta r_t}$

> 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.

- $List_t = \beta_0 + \beta_P lnPermits_{t-1} + \beta_{Int} lnInterest_{t-1} + \beta_{Inf} lnInflation_{t-1} + r_t$
- $\Delta List_t = \underline{\beta_0} + \beta_P lnPermits_{t-1} + \beta_{Int} lnInterest_{t-1} + \beta_{Inf} lnInflation_{t-1} + \rho(List_{t-1} - \beta_0 - \beta_P lnPermits_{t-1} - \beta_{Int} lnInterest_{t-1} - \beta_{Inf} lnInflation_{t-1}) + \varepsilon_t$
- $\underline{\Delta List_t = \beta_0 + \beta_P lnPermits_{t-1} + \beta_{Int} lnInterest_{t-1} + \beta_{Inf} lnInflation_{t-1} + \rho(List_{t-1} - \beta_0 - \beta_P lnPermits_{t-1} - \beta_{Int} lnInterest_{t-1} - \beta_{Inf} lnInflation_{t-1}) + \varepsilon_t}$
- That second equation can then be distributed out and further reduced.

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

- In order the functions do as follows, renaming the *date* variable to *datestring*, generating a new variable called *datec* that is the *datestring* variable in YMD format, generating another variable, *date* that is *datec* in monthly data form, then formatting the *date* variable as a timeseries data-type, then setting the beginning of the time-series data, and finally generating a new *month* variable that is the month of the date.
- We do all of this to make sure that our data is in a format that STATA can work with and so that we have all of the variables we will need later ready to go at the beginning of the analysis.
- "<u>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.</u>"

> 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 *ac* command generates an autocorrelogram graph while the *pac* chart generates a partial autocorrelogram. *dfuller* runs a Dickey-Fuller test on the data. The Dickey-Fuller test has an option for *lag(12)* which lets us lag for 12 months (an entire year) of data.
- AC and PAC graphs can be used in conjunction to identify ARIMA models. If a point is significant, it extends beyond the shaded boundary. For the non-differenced models, we see that significance decreases as time passes. We can then look to the PAC chart and see that there is significant correlation in the first lag and correlations that are not significant. This suggests a higher order autorgressive term. In the differenced models, the AC graph spikes are right at the edges of the range. If they are significant, it suggests an autoregressive term, if they are not significant, it suggests a moving average term. The same goes for the PAC but instead the insignificant values suggest autoregressive while the significant

suggest moving average.

- The Dickey-Fuller test is interpreted by its p-value. I don't remember what it does and everything I have says that it tests to see if there is a unit root but I have no clue what a unit root is and can't figure it out. I found this article https://stats.stackexchange.com/questions/29121/intuitive-explanation-of-unit-root which has a very funny joke at the bottom of the accepted answer. If I'm understanding what A.A. Milne 2.0 is saying, the model does not have a unit root because the p-value is less than one, the data will converge back to the same spot.
- "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."

> 5. Four sets of models are estimates. 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?

- Model 1 is only lagged, model 2 is lagged with the date and month indicators, model three is lagged and differenced, model 4 is lagged and differenced with the date and month indicators.
- It's possible that these are AR/DL models in all four forms. None, autoregressive, distributed lag, and autoregressive distributed lag
- I'm going to choose Model Set 3. There's something bothering me suggesting maybe I should choose 4 instead but I'm going to stick with 3. I'm not super confident with why, but I feel okay with it and I'm running out of time. It's the only one where there's ever any enough evidence to reject the null hypothesis of the Breusch-Godfrey test.
- "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?

- The first set of models is estimated only using the first lag. The second set of models uses the first and second lags.
- The difference does matter because it changes how far back the model looks when making its predictions.
- I think the first option is better because it is taking all three variables into account rather than just two. Like I say below, it is best to include all hypothesis variables when testing.
- "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.

- (I'm not entirely sure if *testparm* or *estat bgodfrey* is the F-test and I don't have enough time to figure it out. I'm going to assume it's *testparm*. I also don't know what *test* is either...)
  Although the p-value is very close to .05, it is still just above it at .0566. This suggests that we our model does not fit the data as well as it could.
- "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?

- One model is like the original model given, one of them is like the first difference model from problem 1, and one of them is like the autoregressive dynamically complete model from question 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?

- You must assume that all relevant variables are accounted for and that there are no *omitted variables*. You must also assume that there is no *multicollinearity* in the data. The first means you shouldn't leave out important data and the second means you shouldn't include two pieces of data that are correlated with each other such as the amount of cereal consumed and the amount of milk consumed. Most people consume those items together, so using them both can be redundant and detrimental to the model.
- "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?

- I think model 3.3 is the best predictor but 3.2 also looks pretty good. The best for testing the hypothesis is 3.2 because it is the only one that takes the number of building permits, the interest rate, and the inflation rate into account. They could be different because when testing a hypothesis, it is important to test all of the variables discussed in your hypothesis. If I say "high consumption of pizza and beers leads to heart disease", I can't only test if pizza leads to heart disease, I have to test both. That said, it might turn out that pizza is a much better indicator than beers and that beers doesn't add much. In that case, the pizza only model would be better because there is less room for error.

- "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."