Problem Set 4: Forecasting and Non-stationarity

1) Forecasting using more variables

1. Estimate an AR(1) model for log(CO2) up until 1990 and use a 1-step forecast to generate forecasts for log(CO2) from 1991-2010. What is the root-mean-squared forecast error?

2. Use the Kuznets curve model you have estimated previously (see below in (1)) and forecast log(CO2) from 1991 until 2010 using 1-step forecasts and contemporaneous RGDP values. Comment on the results (e.g. RMSE) and compare them to the simple AR(1) model estimates. The model to be estimated:

$$\log(\text{CO2})_t = \beta_1 + \beta_2 \log(\text{RGDP})_t + \beta_3 \log(\text{RGDP})_t^2 + \epsilon_t \quad (1)$$

   Note: these are not "true" forecasts as we are predicting values in time t using RGDP information from the same time t.

3. The partial auto-correlation function suggests that we should include one lag of log(CO2) into the model. Estimate the model by adding a lagged value of log(CO2) and compare the forecasts over the same time period to the above models:

$$\log(\text{CO2})_t = \beta_1 + \beta_2 \log(\text{CO2})_{t-1} + \beta_3 \log(\text{RGDP})_t + \beta_4 \log(\text{RGDP})_t^2 + \epsilon_t \quad (2)$$

Solution:

1. The auto-regressive model forecasts are the same as in problem set 3. The root-mean squared error of the AR(1) model forecast is equal to 0.031.

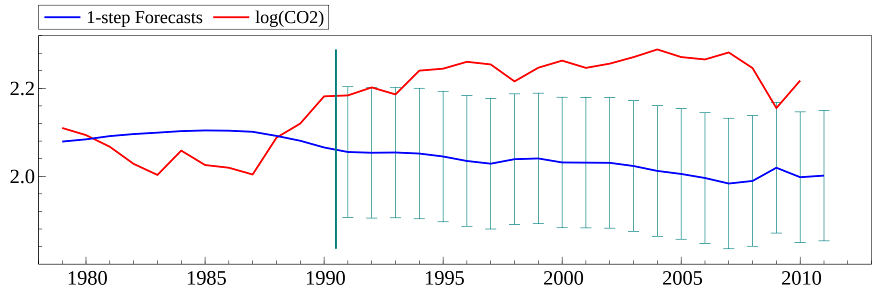
2. The estimated model is:

$$\log(\text{CO2}) = \underset{(7.41)}{-\ 71.9} + \underset{(1.55)}{14.7} \log(\text{RGDP})_t - \underset{(0.0811)}{0.73} \log(\text{RGDP})^2{}_t$$

   with the diagnostic tests given by:

```
AR 1-2 test:      F(2,25)   =   14.855 [0.0001]**
ARCH 1-1 test:    F(1,28)   =  0.51398 [0.4794]
Normality test:   Chi^2(2)  =   2.7697 [0.2504]
Hetero test:      F(3,26)   =  0.43961 [0.7266]
Hetero-X test:    F(4,25)   =  0.83241 [0.5174]
RESET23 test:     F(2,25)   =   9.0325 [0.0011]**
```

   Forcasts are shown in the Figure below:

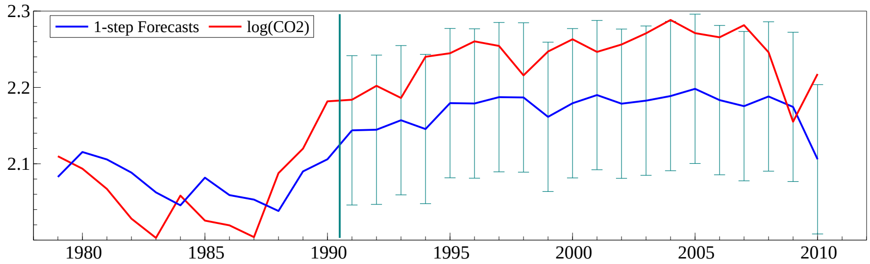


First, note that while we are "forecasting" log(CO2) in terms of model timing these are not true forecasts - the model uses the contemporaneous values of log(RGDP) to generate the forecasted values. As an example, consider the forecasted log(CO2) value in 1991, it is estimated using the model: $\log\widehat{(CO2)}_{t=1991} = \hat{\beta}_1 + \hat{\beta}_2 log(RGPD)_{t=1991} + \hat{\beta}_3 log(RGDP)^2_{t=1991}$, thus the forecast for 1991 uses information on other variables from the year 1991. Despite the use of this contemporaneous data, the model forecast performance is poor. The forecast RMSE is 0.22, vastly higher than in the auto-regressive model used before. The model systematically under-predicts log(CO2) emissions over the entire forecast period. This is likely driven by both model mis-specification (the model fails the test for residual auto-correlation and functional form specification), as well as likely structural change that would have resulted in different parameter estimates had the whole sample period been used (compare the estimated cofficients of this model to the one estimated on previous days). Comparing the forecast performance to the AR(1) model, the auto-regressive model, despite not using contemporaneous information and omitting log(RGDP) from its formulation provides a much closer forecast fit.

3. Including an auto-regressive term we estimate the Kuznets curve model as:

$$\begin{aligned}\log(\mathrm{CO2}) \quad = \quad & \underset{(0.112)}{0.673}\ \log(\mathrm{CO2})_{t-1} - \underset{(9.85)}{20.4} + \underset{(2.02)}{4.22}\ \log(\mathrm{GDP})_t \\ & - \underset{(0.101)}{0.211}\ \log(\mathrm{GDP})^2_t\end{aligned}$$

with a forecast RMSE of 0.075. Forecasts are plotted here in the Figure below:



While the inclusion of an auto-regressive term improves the forecast performance relative to the static Kuznets model, the best forecast performance is still achieved by the simple AR(1) model. Even though the forecast performance is better in the simple AR(1) approach, the simple auto-regressive model, however, does provide less insight into the underlying processes as we are only modelling the series using past values of itself which reveals little about what the underlying drivers are. Forecast performance and model 'usefulness' are often not closely linked.

2) Linear Trend

Think back on the graphs of the series we are modelling from problem set 1. As visual inspection shows, the log(CO2 Emissions per capita), log(RGDP), and $(\log(\text{RGDP}))^2$ are slightly trending. We now investigate these underlying trends.

1. Investigate the presence of a linear trend in the series: Augment the previous model with a linear time trend. Is the coefficient on the linear trend variable significant? What does the result tell you about whether a trend should be included or not?

Solution:

1. Adding a linear trend to the previous model yields the following results:

$$\log(\text{CO2}) = \underset{(7.51)}{-\,51.4} + \underset{(1.62)}{10.4}\ \log(\text{RGDP})_t - \underset{(0.0877)}{0.505}\ \log(\text{RGDP})^2{}_t + \underset{(0.00394)}{0.00256}\ \text{Trend}_t$$

Using a simple t-test for the coefficient on the linear trend variable we find the t-ratio to be 0.649, with an associated p-value of 0.52. The trend variable is not statistically different from zero, we thus conclude that a linear trend should likely not be included in the model specification.

4) Big Picture

These questions have no clear right or wrong answers, but rather require your interpretation and creativity.

1. Overall, given the evidence from your estimated models, what do you conclude on the the environmental Kuznets curve for Japan?
2. How could the models you have estimated be improved further?
3. How would you structure an analysis of the environmental Kuznets curve in a short paper, from presenting the data to analysis and conclusions (providing a short outline is sufficient)?

Solution:

1. The static models of the environmental Kuznets curve so far suggest some evidence of the theory that per capita CO2 emissions increase with GDP per capita with a decreasing marginal effect. The coefficients in the estimation were statistically significant and suggest a turning point, i.e. decline in emissions per capita given GDP per capita around an income level of approximately 34500 USD (in 2011 dollars) for Japan. This value was only recently reached and it remains to be seen whether Japan does follow the environmental Kuznets curve to the extent of leading to a decline in per capita emissions. The conclusions should be interpreted with care, given that the models were mis-specified as we were able to see from the misspecification tests.
2. The model used is a highly simplified abstraction from reality. Given the misspecification tests further lags of the dependent variable could be included or a different functional form chosen. Equally, further explanatory factors should be explored, such as the emission intensity of industries, economic growth (rather than the level) and other potential variables explaining per-capita CO2 emissions. Inspirations for which variables could be included can be drawn from the literature on the environmental Kuznets curve. As seen in the lectures, an automatic model selection approach could be considered by starting with a very general model and using model selection to reduce it to a specific one.

3. There are many ways an analysis of the environmental Kuzents curve for Japan could be structured, an example is provided here:

Example structure of a paper on the environmental Kuznets curve:

1 Introduction

The introduction should outline your overall idea and what theory you are testing. Describe the theory behind the environmental Kuznets curve, cite related literature and provide a short overview of what other researchers have found in this field. Also provide a short overview what your further analysis will be about.

2 Data

The data section should introduce all the data sources including references to where they were obtained from, the time frame they cover, and provide summary statistics (mean, standard deviations, units) as well as plots of the series in levels and first differences.

3 Methodology

The section on methodology should outline the estimation procedure and how the estimated model links to the theory. In the case of the environmental Kuzents curve, you should specify here that the estimated model will include real GDP in levels and as squared term to capture potential diminishing marginal effects. You should also specify how you will determine the appropriate lag lengths of the model. You can do this by plotting the (partial) auto-correlation functions. Further you can discuss whether you will be using forecasts, models in differences, or automatic model selection.

4 Results

The results section should provide a clear overview of your estimated models. Ideally the regression results are reported in a table such as given here:

Table 1: Modelled Variable: log(CO2 Emissions per capita)

Variable	Estimated Coefficient
Constant	-46.48 (4.15)**
L(RGDP)	9.32 (0.85)**
$L(RGDP)^2$	-0.45 (0.044)**
R^2	0.96
T	51
AR 1-2 Test	67.647 [0.00]**
ARCH 1-1 Test	48.077 [0.00]**
Normality Test	1.3641 [0.50]
Hetero Test	5.6484 [0.002]**
RESET Test	20.552 [0.00]**

It is important to include the estimated coefficients, standard errors, number of observations, R^2, and results of model mis-specification tests. You can estimate multiple models e.g. in first differences, including or excluding various variables but make sure you report your results in full and in a neatly organised manner. Be sure to include figures of your fitted values and model residuals.

5 Discussion

The discussion should put your results in context of the theory your are investigating. Do the results support the environmental Kuzents curve? Do they reject it? What does this mean in the broader context and how could the model be improved further?

6 Conclusion

The conclusion should concisely summarize the points raised in your analysis without adding new material.

7 References

All literature cited (directly and indirectly), together with all data sources used should be listed in the references.