Bo Sjo
2020-10-29 (EViews & PcGive, can also be done in gretl)

Exercise 2

Intervention Analysis – How to use Dummy Variables, Seasonal effects, why to use logs, and avoid spurious de-trending.

A small part of this exercise requires that you have read about how to identify AR(p) and MA(q) processes with autocorrelation and partial autocorrelation functions.

1. Introduction

Intervention analysis is just another word for identifying and dealing extreme observations and shift in the data.

Modelling economic time series involves checking and transformation of data. Here is a collection of small exercises that deals with practical work related to economic time series.

The following topics are covered:

- Taking logs
- Detecting outliers/extreme vales
- Impulse dummy
- Step dummies
- Seasonality
- Spurious detrending

When working with time series you usually always take log of variables. You must always look for outliers, extreme observations, among the data. There could be miscoding or simply extreme events. If there is only a few or two extreme event there is not enough information to model such phenomena.

2. Logs

When working with economic time series it is usually required that you work with the natural log transformation of all non-stationary variables. Taking the log of a non-stationary variable is a way of making the variance of the series more uniform across the sample. Changes in a log variable come close to looking at percentage differences.

In finance and economics log differences (natural log) is always preferred over percent because the former corresponds to continuous compounding, (and that percent return are subject to Jensen's inequality[1]).

We have

$$\Delta \ln x_t = \ln x_t - \ln x_{t-1} \approx \frac{x_t - x_{t-1}}{x_{t-1}}$$

When taking logs of stock prices (adding dividends to the price series if you want) you get continuously compounded return. This is what you should use in all regression and correlation calculations.

To see how logs improve things look at the quarterly GDP series of Mozambique in the file **"mozam_gdp_q.xls".** Compare the series before and after taking the log. Before the log transformation, the series grows in an exponential way and the volatility increases. After it has been logged it grows in a more linear way and the variance is more uniform across the sample. Every change, at any time, is now a percentage change (close approximation) and not a change in the local currency unit. Any regression model will give better fit and inference and interpretation is easier.

[1] Jensen's inequality (or paradox) is E(1/x) > 1/E(x). See also the web.

Another example is the Industrial Production Index (IPI) for Sweden during 1913-2010, stored in **"swe_ipi_year.xls".** Graph the series before and after logs. Sweden is affected by the world wide recession in 2007-10. The question is how much is Sweden affected in an historical perspective? Compare the series before and after logs to see how important the recession really was.

Swedish GDP is also affected by seasonal effects. One way of dealing with seasonal is to look at take the fourth difference of the series $\Delta_4 \ln x = \ln x_t - \ln x_{t-4}$. This will make comparisons over time easier. You can also take the first log difference and rank the series from large to small.

3. Info about Intervention analysis - Identifying extreme observations – A long intro

In time series data extreme observations are common both in the data series as well as in estimated residuals. The latter is quite crucial because there are situations when we need normally distributed white noise residuals, and also make sure that extreme individual observations are not causing your results. In statistics the general term for identifying extreme value (outliers) is intervention analysis.

Here are two examples of dummy variables. The first example deals with one extreme value, which calls for an impulse dummy. The second example deals with level shits in the data and calls for the introduction of several shift dummies.

Some time series display extreme observations, or there might be shifts that move the mean of the series up or down during the sample. In an ideal world we should be able to model such events, but in the real world we are often forced to leave them unexplained, or treat them as some exogenous events. Events such as major strikes, devaluations, tax (VAT) changes, war etc. can and often will cause data to change.

In some situations we will adjust for changes in a data series, most often in the dependent variable in a regression. In other cases, the adjustment is made after looking at the residuals of an estimated model. By imposing a dummy for an extreme value in the residual, the model will usually fit better the data better and pass misspecification tests. Dummies of all kinds can be used in ARMA models as well as in other types of regression models. However, shift dummies will affect tests for cointegration in such a way that critical values change.

To accurately model series with extreme observations, or shifts in the mean, it might be necessary to 'remove' or 'dummy out' the extreme effects. The analysis which leads to identifying extreme observations in time series is in statistics called intervention analysis, in econometrics it is common to simply refer to dummy variables. In time series we talk about two types of dummy variables; impulse dummies or step dummies. Both are explained below. In addition, there are also seasonal effects, which can also be handled with seasonal (impulse) dummies.

In a cross-section study, extreme observations can simply be removed from the sample. In a time series setting it is not possible to remove say one month or a year from the series. The fundamental question is on what grounds do we identify and classify some values as 'extreme'? The answer is that whatever we do it might be *ad hoc*. A strike might, or a devaluation, be one single event, or it could be something that happens repeatedly on the labour market that we study. We might simple have a short sample, or we should try to model the mechanisms that lead to a strike, or devaluation. But, if there only a few observations of events like strikes and devaluations, we will might not have enough information to model them.

On the other hand, if we do not remove extreme effects our analysis of trends and dynamics might go quite wrong. One, very common motive for removing extreme outliers, is that our modeling techniques and in particular inference builds on the assumption of normal distributions. For instance, to do a standard significance test in an OLS regression, we should have white noise normally distributed residuals. Thus, a test for normal distribution of the estimated residuals should not be rejected. However, tests for normal distributions are sensitive to small sample sizes and extreme values. By removing the extreme values it might be possible to claim (or not reject the null) that a series, like estimated residuals, is normally distributed.

In this exercise we look at three examples of intervention analysis; one extreme value, a few extreme values that affects the assumption stationarity, and some extreme seasonal effects that might shift over time.

In the end, as in most of time series modeling, everything becomes a matter of judgment and thus your ability to justify your decision.

4. How can an outlier be detected?

To identify outliers you will have to use two methods, you can look at the graph of the variable and look at the estimated residual from a regression of the variable. The important outliers are those that show up in the residuals of your estimated model.

Run the regression, look at the standardized residuals and look for the extreme values. You will typically see the extreme values, structural breaks etc. Standardized residuals are the estimated residuals dived by their standard errors.

In general, extreme values are typically standardized values above +/- 3.5% around the mean. If necessary to build a model with normally distributed residuals, this value can reduced step by step to +/- 3% and to +/- 2.5%.[2]

From the graphs of the standardized residuals you can identify the extreme values.

To identify potential outliers, modeling, you can always run a simple OLS regression such

$$y_t = a_0 + e_t ,$$

and analyse the residual.

The constant will capture the sample mean of y_t and the residual will be the mean adjusted y_t series. The standardized residual of this regression (the mean-adjusted y) will reveal extreme values, as well as seasonal effects etc.

EViews

[2] Notice that gretl uses a more narrow band. The program tests if the standardised and normalised residual is significantly different from zero. (You can transpose an estimated residual to N(0,1) by dividing each estimated residual with the estimated standard deviation of the residuals.

In EViews look under View/Actual, fitted, residual/Standardized residual. The Actual Fitted Residual graph can also be helpful. You can also look at the residuals in a spread sheet in which can also sort the residuals in terms of size.

PcGive & gretl:
See info below

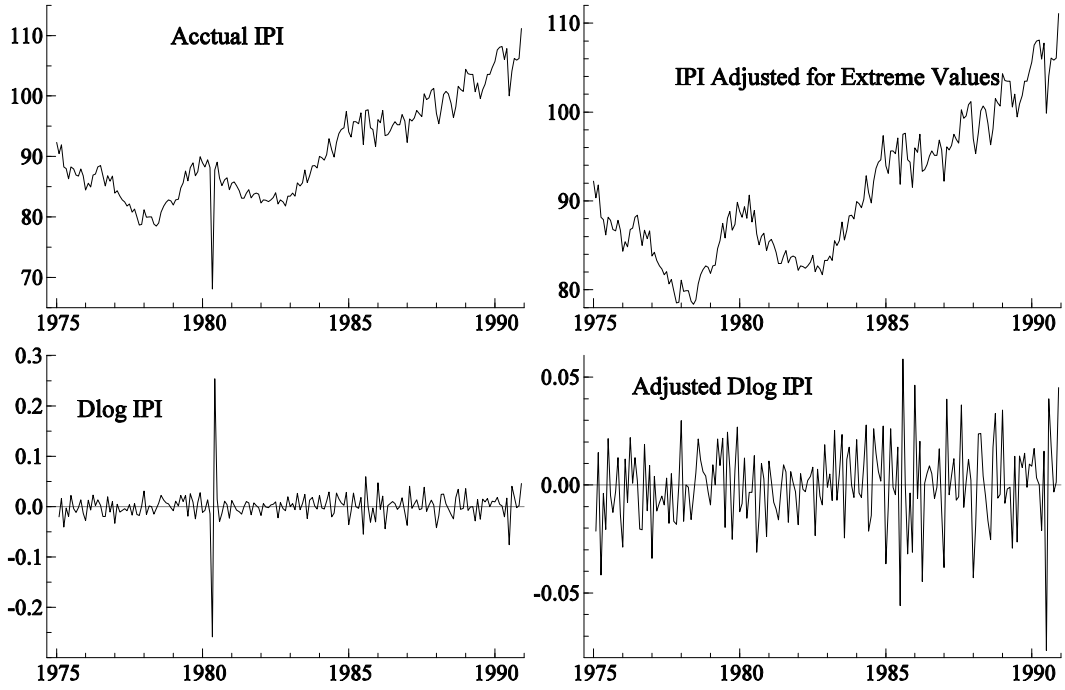
5. A Single Extreme Value – Impulse Dummy Data: swe_ipi_mon.xls

Load the series "**swe_ipi_mon.xls".** This series is the monthly industrial production index for Sweden, and it contains one extreme value. Start with graphing the variable (you can take logs if you want), in level and in first difference. Do you see the extreme value? Try to remove the extreme value with a suitable dummy variable.

Start by running an OLS regression of the series in levels against a constant. Look at the estimated residual. Notice that the residual is simply the original series adjusted for its mean. You can confirm this by adding the original series to the graph and adjust the graph for different means.

Identify the period for the extreme value. Create an impulse dummy for this period. An impulse dummy takes the value (0,0,0,0,..,1, 0,0) where unity is imposed at the period of the extreme observation.

The following graph shows the series before and after adjustment:



Notice that I removed the outlier with OLS and put back (added) the mean in the graph at the upper left.

Next, take log difference, meaning that you remove the trend, and then remove the extreme values, using only one dummy variable.

If the chosen dummy (Dum) actually removes the outlier is seen by looking at the residual of the regression,

$$y_t = a_0 + B*Dum + e_t .$$

In this regression, if you identify the right period, the dummy should be highly significant, and the residual should display no extreme value. The effect of the impulse dummy is to allow for a different mean in the series for this particular observation.

The conclusion is that by adding an impulse dummy, an extreme value can be 'removed' from the data.

Next try to remove the extreme value from the (log) differenced series. Use the first difference operator. After differencing, you get two extreme values of opposite sign but very similar in absolute values. Now put in an impulse dummy of the type (0,0,0,0,..,1, -1, 0,0). The order of +1 and -1 does not matter. Next run a regression with constant and dummy variable, the extreme values should be gone if you check the residual (the mean-adjusted series). You can say that you filtered out the extreme values.

Facts: This extreme event is a rare labour market conflict in Sweden. Such conflicts are very rare in Sweden, so in this case it seems highly motivated to remove the observation from the data. An impulse dummy is of the type (0,0,0,1,0,0…) will do the trick.

Shift dummies are used when we need to compensate for shifts in the mean of the data over longer periods, type (0,0,0,1,1,1,1…)

EViews:

See EViews Introduction (my memo updated) for how to create dummies. In EViews use the View /Actual fitted, residual window to locate outliers. In a graph window moving the cursor over the outlier will inform about the data (or number) of the observation.

PcGive:

In PcGive use the Calculator to create the dummy variable. In the graph window use the cursor to see the data of the outlier. And to check, in the Test menu, under Further output, find the option to Print large residuals.

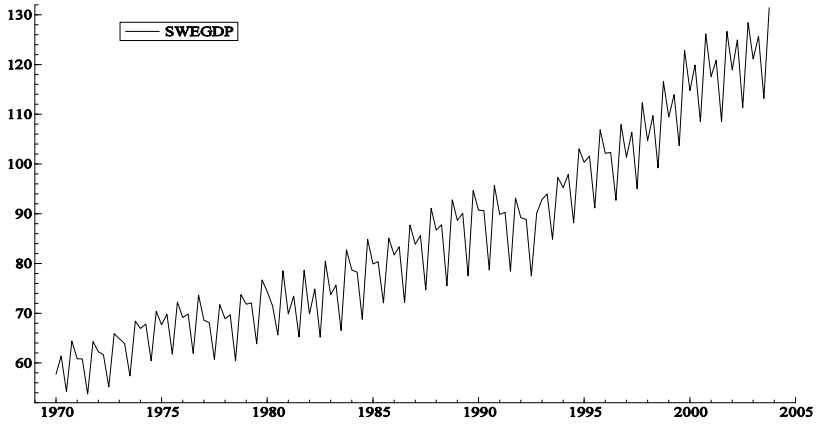
Gretl:

After regression look under test and influential data. Gretl indicates large residuals. Create dummies under Add and Observation range dummy.

6. Seasonal Effects – Seasonal Dummies or X12

Seasonality can be dealt with by (1) imposing seasonal dummies, (2) use seasonal removal estimation program or (3) use seasonal differencing. For EViews se my EViews intro. PcGive and gretl is let you add seasonal dummies directly since it is preprogrammed routines. The last alternative is X12 (or the recent version X13/X14). This is an automatic (black box) routine for removing seasonality. Chose standard pre-set values for now.

For this example, load **"swe_gdp.xls**", this is the real Swedish GDP (index) which displays strong seasonal effects, as the following series shows,



This series have very problematic seasonalities.[3]

In addition the series contains three different segments during which the calculation of GDP has changed. The seasonality is also changing. Try to deseasonalize the data with seasonal dummies.

Try, X12 (orX13/X14), and compare. Compare the outcome?

In X12 try both to deseasonalise the level and the first difference.

EViews:

Create dummies according to instructions.
To use X12, from Workfile open the series in a separate window. Click on Proc and the Seasonal adjustment. Chose Census X12. Use the default options as far as possible.

PcGive:

Add dummies from the estimate model menu. To use X12: In version 12 of PcGive, X12 hides under the "Run" menu. In X12 you are looking for the series called D11, finally adjusted series. To save the adjusted series go to the Test menu in X12 and save the series D11.

[3] SCB hired the world's best experts in sesonality to find out how to model this seasonality, but they couldn't do it.

7. Spurious detrending

In the old days, prior to the end of the 1980s, it was common to remove trends from macro data by regressing the variable against a deterministic time trend. It meant that trends were assumed to be deterministic trends that could be removed by conditioning on a deterministic trend line. It doesn't work well.

Trying to remove a deterministic trend out of a data series that has no deterministic trend can have serious consequences. If a series has a stochastic trend (the series is integrated) it is not possible to detrend that variable by regressing it against a linear deterministic trend. In fact the outcome will be wrong and we talk about spurious de-trending. The spurious de-trending will introduce a correlation structure in the data.

The idea was to regress the trending variable against a linear trend,

$$y_t = \text{const} + B_1 * (\text{trend}) + y_t^+,$$

where trend = 1,2,3,4,…..

The residual in this regression (y_t^+) will be the detrended and stationary if the series follows a determinist trend.

To see how it works, or rather not work, open the data rw.xls, and run a simple trend regression as above. Save the residual.

Next, load the seasonally adjusted GDP form the source above and take logs of the GDP series.

Look at both series in logs and in first differences. The series are clearly integrated in levels and possibly integrated in first difference as well.

Try to detrend the series with a linear time trend, and look at a graph of the residual y_t^+ . Does it look detrended? Does it predict the future well?

Next, try a better fit, add quadratic linear trend to the regression$_1$

$$y_t = \text{const} + B_1 * (\text{trend}) + B_2 * (\text{trend}*\text{trend}) + y_t^+$$

This will show a better fit that the former series. Again, does it look detrended? Does it predict the future well? The answer might be yes, and then perhaps no especially if you compare with first log differences.

If you look at the autocorrelation functions, and the partial autocorrelation functions, you can see that the stochastic trend is still in the series. In addition, observe that the number of times the series cross its mean axis are quite few.

If the series is integrated without a linear trend the outcome is what is called spurious detrending.

To see this how works, take the series trend.xls, regress it against a constant and a trend, and inspect the residual with graph and time series plots PACF (Partial Autocorrelation Function)

and ACF (Auto Correlation Function) . The pattern you see there is a pattern you introduced by regressing this integrated variable against a linear trend. This series is not detrended, it only has an irrelevant time series structure imposed by wrongly removing a deterministic trend that wasn't there in the first place.

If you regress the stock.xls series or the GDP series in the same way you will se a similar pattern in the ACF:s of the residuals or the "detrended" variables.

How to remove the trend in GDP data. Answer: typically the dominating trend is a stochastic trend that is removed by first (log) differencing.

8. Shifts in the Mean – Step Dummy Data: swe_exch.xls

This is a very complex exercise, but it is worth a try to see far we can come with linear step dummies, at least during the "basket period until the early 1990s. A step dummy is of the type (0,0,1,1,1,1,0,0,…), and shifts the mean (the constant) over a specified part of the sample. This can be done for one more segments of periods over the sample period.

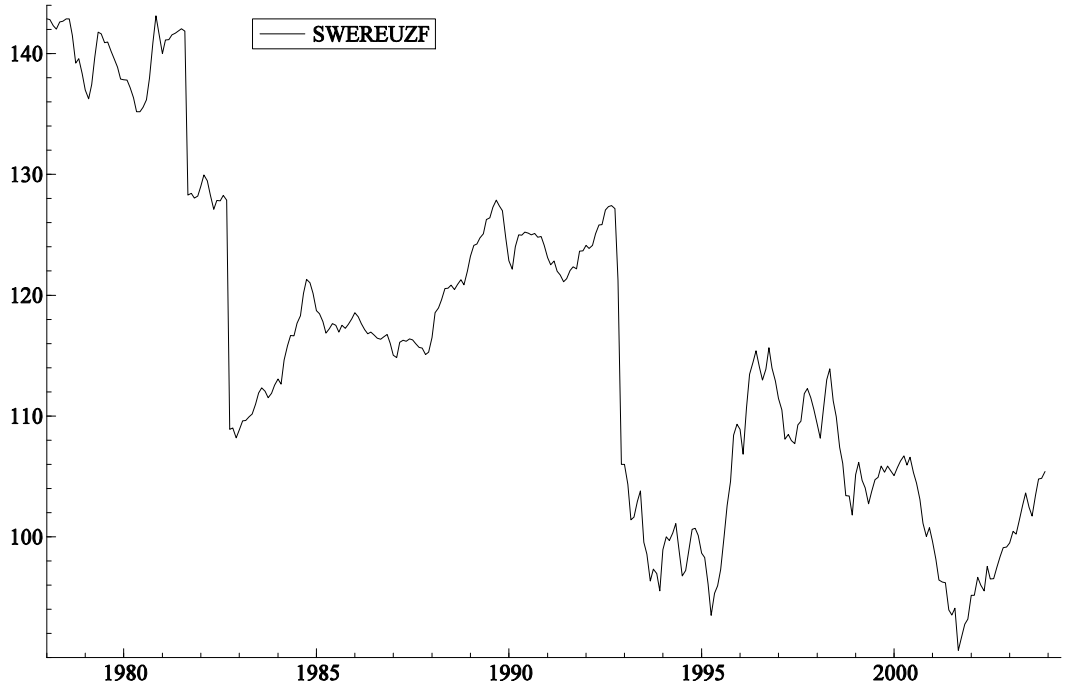
For this example, load the data set "**swe_exch.xls**", which is the Swedish exchange rate index. In the graph, identify the devaluations, put in step dummies to make the series stationary in the sense that the mean becomes the same over the sample.

At a first glance, the series could be labeled as non-stationary, since the mean of the series is changing. But what type of non-stationary is it? Is it a trend? Is it an integrated variable? The non-stationarity comes in the form of shifts in the level of the data associated with devaluations, which looks more like segmented shifts in the data.

As a background, Sweden tried to keep the value of Swedish krona stable against a basket of currencies during 1997 up to 1991. Within the basket the krona was allowed to move up and down around the target +/- a few per cent. Before 1977 Sweden was in the so-called currency snake and tried to keep the value against the krona fixed towards the German mark. There is one devaluation of the krona in 1977, another in the fall of 1981 and a third devaluation in the fall of 1982. In the 1990 the krona is left to float freely. Try to find the exact month for these devaluations and put in step dummies for these periods.

The following graph displays the Swedish exchange rate according to the trade weighted index. Identify when the krona was left to float. Put in step dummies for the period 1977-199x to make the series at least relatively stationary during this interval. You cannot and should not do anything with the floating period.

In principle, step dummies show produce something that look stationary. However, it seems that the exchange rate behaves in a non-stationary way within the bands. Notice, even though the moments are bounded it will still appear to quite be non-stationary within the bands. That is could be a consequence of the exchange rates being like asset prices with free trading withing the band. Under all circumstances, the devaluations create jumps in the data that is hard to model except for imposing dummies. The size and timing of the devaluations are the outcome of policy and political decisions. And we only have a handful of observations. Too few to make a model.



This series display clear shifts in its mean. (Compare graphs in levels and in first differences). The series is clearly stationary after differencing. However, it is also an integrated variable, or is it simply a stationary variable that moves with segmented deterministic trends?

Yu can run regression with a constant only, and later think about including step dummies for all sub-periods.

Start by graphing, level and differences. Look at the PACF and ACF in a graphics window. Use the information to ask the simple question if the series is integrated or not, should it be differenced to become stationary?

The answer is that there might be a different approach to differencing to achieve stationary for this series. The series could be modeled with step dummies instead, and by stationary around these steps. There are some clear shift in the series, identify these shifts and construct step dummies for them.

In the end you can filter-out the shifts in the mean and construct a stationary data series. From the data you are able to identify the exact dates of every devaluation. In all there are 3 devaluations before the krona is allowed to float freely in the early 1990s.

Run the regression against the step dummies and look at the residual. The residual is the filtered series, which looks more stationary than the original series.

This series can mistakenly be assumed to be integrated of order one I(1) between 1977 and 199x.

If you are too mechanic when it comes to differencing series you might miss the fact that series is stationary between the jumps.

Facts: For a large part of the period the exchange rate is partly fixed. It is a so-called basket system. The exchange rate system is an index which is kept within a band. Since the exchange rate is not allowed to move outside the band it is in effect stationary as long as the target index rate is not changes. After 1992, the exchange rate becomes freely floating. There are two devaluations and one depreciation in the sample.

Remember, that both ACF and PACF are sensitive to shifts in the mean. The lag length will look much longer than it really is. To compensate for the shift the ARMA model must add more lags to mop up all autocorrelation. (The same holds for so-called unit root tests like the Dickey-Fuller test, this test is biased towards accepting I(1) when there are segmented shifts in the data. Run an ADF test on the data series to verify this.)

This exercise is a bit problematic, but in principle it should be possible to construct a (quite) stationary series after conditioning on shift dummies for the different in-between devaluation periods, since the exchange rate is bounded between devaluations.