Ekonomisk statistik
Economic statistics

Master course

Daniel Thorburn

Autums semester 2010

Stockholm University

3.1 Time series - Linkage

Daniel Thorburn

Ekonomic statistics

Autumn semester 2010

(not included in this version)

3.2 Time series – Seasonal adjustment

Daniel Thorburn

Ekonomic statistics

Autumn semester 2010

Partition into Components

- The series Y = T + C + S + K + E
  - T, ”Trend” is the development in the long run (What is meant by long run differs, ususally one year or more)
  - C ”Cycles” (Transient deviations from the trend which are not due to the season nor limited to one measurement (the economic climate, booms and recessions) +
  - S ”Season” (E.g. summer, christmas, school year, vacations, salary day within month …) +
  - K ”Calendar effects (length of month, Easter time, number of working days …) +
  - E ”Random Error” (effects vvalid only one measurement like measurement errors, weather, lockouts, but sometimes also accrual effects)

Transformations

- Additive – multiplicative – seasonal adjustment
- Economic variables are usually handled best as multiplicative, i.e. take the logarithm before analysing.
- Other transformations like Box-Cox transformations exist and can be used if you have a good reason
- One disadvantage, though, is that you sometimes want additivity (Trade balance should equal exports – imports)

Simple methods

- Comparison with the same period last year/month/period (The Swedish economy (quarterly GDP) has gone up with 4.5 % since last year)
- Traditional
  - Trend* = Moving average (e.g. the sum of the last period (12 months))
  - Season* = (weighted) average for some (e.g.5) of the last observed values for the same season
  - Error* = Observed value - Trend* - Season*
- Seasonal adjusted series is sometimes Trend* and sometimes Y – Error*

A simple example: Model Y = T + S + E

- Data (quarterly data)
  1 1 2 3 4 3 4 7 5 3 5 5 5 8 8 12 11 10 9 11
- Make a 4-quarterly average ((½*1+1+2+3+½*4)/4=2.1) for T*
  - - 2.1 2.8 3.2 4 4.6 4.8 4.9 4.8 4.5 5.1 … 10.4 - -
- Five year moving average
  - - - - - - - - 5.0 5.2 5.6 7.6 - - - - - - - -
- Take away the mean (an alternative is the quarterly average) S*
  - - - - - - - - -0.6 -0.8 -0.2 1.8 - - - - - - - -
- The error term will be estimated as E*=Y-T*-S*
  - - - - - - - - 0.5 1.2 0.4 -1.6 - - - - - - - -
- This was only one of many possible, but simple methods
- But you can see the problem. This adjustment technique will not work properly for the last two years. Trend was not estimated for the last half-year.
- There are several possible solutions. We will discuss three different ideas very shortly.

Three common approaches

- The exponential smoothing group. Remake all averages to cover only observed values but do it in an intelligent way
- The Arima group. Use Arima techniques to predict the series backwards and forwards and use a simple adjustment technique on the predicted series.
- Dynamic time series models. Make a complete model for the whole series and estimate all components using the model.

The exponential smoothing - group

- Exponential smoothing $Y^*_t = \alpha Y_t + (1-\alpha)\, Y^*_{t-1}$. where α is a number between 0 and 1.

- Why does this work? What α to choose?

  We illustrate a by formulating state space model:
  Innovation: $T_t=T_{t-1}+\eta_t$; Measurement: $Y_t=T_t+\varepsilon_t$.

$$Y^*_t-T_t = \alpha(Y_t -T_t)+(1-\alpha)(Y^*_{t-1} -T_{t-1})-(1-\alpha)(T_t -T_{t-1}) =$$
$$= \alpha\varepsilon_t+(1-\alpha)(Y^*_{t-1} -T_{t-1})-(1-\alpha)\eta_t$$

- Under equilibrium: $Var(Y^*_t-T_t) = (\alpha^2\sigma^2_\varepsilon + (1-\alpha)^2\sigma^2_\eta)/(1-(1-\alpha)^2)$
- This is minimised for $\alpha = \sigma^2_\eta/(\sigma^2_\varepsilon+\sigma^2_\eta)$ (ABLUE, Asymptotically best linear unbiased estimate)

More complicated variants

- Exponential smoothing with a seasons

  $Y^*_t = \alpha(Y_t - S^*_{t-p}) + (1-\alpha)\, Y^*_{t-1}$

  (exponential smoothing for level)

  $S^*_t = \beta\,(Y_t - Y^*_t) + (1-\beta)\, S^*_{t-p}$

  (exponential smoothing for seasons)
- Holt-Winter's method without season

  (exponential smoothing with linear trend)

  $Y^*_t = \alpha Y_t + (1-\alpha)(Y^*_{t-1} + T_{t-1})$

  (Exponential smoothing. T is here the slope of the trend)

  $T^*_t = \beta(Y^*_t - Y^*_{t-1}) + (1-\beta)T^*_{t-1}$

  (exponential smoothing for trend)
- Holt-Winter's method with season …

The Seats-Tramo/X12-Arima-group

(Use ARIMA-models for the series)

- Large and widely used program packages. (Seats-Tramo is created by Maravall from the Spanish National Bank, X12-Arima by Dagum at Statistics, Canada)
- Statistics Sweden uses SEATS-Tramo today
- Mostly used as black boxes
- The principle is as above. Season + Error = Y – Moving average of past and future periods
- But future are unknown.
- Estimate an ARIMA-process and predict future values using it. Use these predicted values as true ones in the moving average.

- The choice of the the ARIMA-model will be extremely important. This is usually done, ad hoc – by looking at the the smoothed series or at the usual and the partial autocorrelationfunctions (ACF and PACF)
- In SEATS/Tramo there is a good automatic black box model for the parameter choices. When handling hundreds of time series like all branches in NACE at the two-digit-level, you must use an automatic procedure.
- The automatic version in X12-Arima is not quite as good according to a test at Statistics Sweden.
- X12-Arima is more traditional Arima modelling (in the way you meay have learnt in a time series course), while SEATS-Tramo uses spectral density in the model identification phase.

Dynamic models – (more like standard theoretical statistics)

- Build a model for everything, i.e. also for the latent parameters. (like we did when we illustrated exponential smoothing but more elaborate)
- E.g.
  - Seasonal pattern: $S_{tp} = S_{tp-p} + \varepsilon_{stp}$ $\quad Cov(\varepsilon_{stp}) = \Sigma \quad$ där $\Sigma_i \sigma_{ijp} = 0$
    (Here S_{tp} is a vector with all p components for the time period tp)
  - Trend $\Delta T_t = \Delta T_{t-1} + \varepsilon_{\Delta Tt}$ (Here t is a number)
  - ”Cycle” $C_t = \beta C_{t-1} + (1-\beta) C_{t-2} + \varepsilon_{\theta t}$
    (A stationary AR(2)), which will be periodic for some parameters
  - $Y = C_t + T_t + S_t + \varepsilon_{Yt}$
- Estimate the parameters. Compute the BLUE predictors of all interesting components in the past and in the future

Direct – indirect adjustment

- You have often many components e.g. in the sum BNP = Σ(all sectors). Should you
  - Adjust the total first and adjust the other components afterwards with the restriction on their sum (direct adjustment)
  - Or adjust every copmonent first and then sum the adjusted components to make the adjusted total (indirect seasonal adjustment)
- There is no general agreement on the answer.
  - Those using simple smoothing techniques usually favour direct adjustment
  - Those in favour of ARIMA-models usually favour indirect if the time series models (sometimes degrees) are similar but th**e**y favour direct if the models are completely different
  - Modellers adjust the series simultaneously using multivariate state space time series models. This gives efficient adjustment if the model is true but to the price of quite complicated calculations and the models may be false
- When the series are adjusted multiplicative the usual advice is to use direct adjustment

Criteria for good adjustment

- It is not easy to decide which adjustment method is the best (except for modelbased where standard rules like MSE can be used
- Three criteria should be mentioned
- Idempotency (if you adjust an already adjusted series nothing should happen). This does not hold for the exponential smoothing group or for simple methods based on averages. But it applies to methods based on dynamic models. The ARIMA-group lies in between
- Small changes when new observations are added. The adjusted value Y^*_t based on $Y_1 \ldots Y_t$ and $Y_1 \ldots Y_{t+1}$ should be similar.

- Small variations in the smoothed series e.g. in $\Sigma(Y^*_t - Y^*_{t-1})^2$ or $a\Sigma(Y^*_t - 2Y^*_{t-1} + Y^*_{t-2})^2 + b\Sigma(Y^*_t - Y_t)^2$ for some constants, a and b, when the effects of S, K and E have been removed in Y^*_t. Or just plot and check.

- Of course one can obtain a smoothed series by minimising this expression. This is called a Hodges-Prescott-filter. For large a it will very smooth almost a straight line and for large b the series will not be smoothed very much.

- When presenting a time series always tell which methods that have been used for seasonal adjustment and smoothing
- Researchers and many economic analysts usually want to use the raw data and do the adjustment themselves
- Smoothing always make the series look nicer and to have a better precision than the true precision.

2.3 Time series – Other aspects

Daniel Thorburn

Ekonomic statistics

Autumn semester 2010

Special problems with time series

- Many time series have a particular structure of the measurement errors. Det kan vid analysen vara bra att känna till det. Annars kan man hitta felaktiga samband med t ex ARIMA-modeller
- E.g. Labour force survey. Every month 7/8 of the sample will be in the sample three months later and 1/8 are fresh. The correlation function for the sampling error will be roughly r(t) =
    - 1 if t=0
    - ~(24-t)/24*r^t i t is divisible by 3 and less than 25 (r varies depending on what variuable you look at, but may for instance be around 0,95
    - 0 if t is not divisiable by 3 or larger than 23
- Standard time series analysis (e.g. Arima) will find a spurious period of three months. (I have heard Swedish economists dicuss what the economic reason for this three months' period is.
- The Consumer price index changes the basket every new year. If you look for that in your analysis you will find it.
- Seasonal adjustment and linkage may also create problems. E.g. a seasonal adjusted series is mostly much smoother than the true series should be. Estimated variances will thus be smaller and the economist will have easier to prove his hypothesis, since the estimated variance is smaller than the true variance.

Longitudinal studies

- Longitudinal studies is a collection term for studies where you follow the same units over time.
- The opposite is cross-sectional studies
- E.g. study the enterprises which have got regional localisation suppoert to see what happens to them and compare to similar firms which did not get support or who got some other type of support.
- Or study firms that grow several years in a row
- Longitudal and crossectional studies have different advantages.
- If you follow persons most people get an increase in the income. This does not mean that the whole population earns more, since their is a group of young persons getting their first job and another group who retires (or dies). The last group has usually higher salaries than the first group. ”The ecological fallacy”

The inspection paradox etc

- Suppose that you want to study the development of firms on the Swedish stock exchange. If you select the 10 biggest and study them for 5 years.
- Then you will probably underestimate the development. Why?

The inspection paradox etc

- Suppose that you want to study the development of firms on the Swedish stock exchange. If you select the 10 largest and study them for 5 years.
- Then you will probably underestimate the development. Why?
- The correct way would be to compare the ten largest with the 10 largest after 5 years. The largest firms can only loose their ranking.

The Inspection Paradox etc

- Select a random sample of in-patients at a certain hospital and ask them about how long their stay at the hospital have been and note when they leave the hospital
- Why will the average of these times be an overestimate of the average time in the hospital?

The Inspection Paradox etc

- Suppose that you want to study the amount of environmental investments during one year. You make a study on a random sample in September and asks many questions on September.
- Those who answered that they did an environmental investment in September are reinterviewed in a special survey and asked about all investments since last September.
- Why will this give a skew estimate of the total environmental investments?

The Inspection Paradox etc

- Select a random sample of all SJ-trains during a certain month and measure how much they were delayed.
- Why will an average of these delays be a misleading measure of how well SJ follows the time table?

The length of the reporting year

- Some companies changes their finanancial year which means that the year will not contain 12 months when the change takes place. E.g. some will have only six months and some 18. There will also be some whose reporting year is July-June; others have Jan-Dec
- This will create problems when making aggregate statistics for a calendar year. It may be tempting that if a firm has ayear of length 12 months the figures from that year will be included in the aggregaqte statistics
- But a firm who changes say reporting years Jul0-Jun1, Jul1-Jun2, Jul2-Dec2, Jan3-Dec3. What should one do the second year when there is two reports?
- It is better to count one half of the first reporting year + one half of the second to make up the figure for the year one. (But the statistics can not be published until much later). Another choice could be to take ony 2/3 of the sum of the two firs reporting years. But then the sum over the years will not be correct.

4.1 Editing – (granskning), correction

Daniel Thorburn

Ekonomic statistics

Autumn semester 2010

(Short version here)

Micro-editing

- 40 % of all data collection costs for enterprises at Statistics Sweden is editing and correction of reported figures
- Micro-editing – Checking of one item and firm at a time. Everything which seems to be suspicious is controlled (E.g. A company having employees but no cost for wages or a unmotivated large change since last year)
- A better approach is to check only items which may be important for the statistics. Ther will probably be more incorrect items in the data-base but the statistics will still be acceptable.

A statistical view on Micro-editing

- You have received a number of responses Z_i and want to estimate $\Sigma_R Y_i / \pi_i$, where Y_i is the value you would have got if you checked that value.
- You are allowed to take a sample S and check, and you have access to auxiliary variables X_i.
- This is a traditional formulation of a sampling problem
- Check unit i with probability p_i (which may be one but not 0) and estimate the total by for instance a difference estimator

$$\Sigma_R Z_i / \pi_i, + \Sigma_S (Y_i - Z_i) / p_i \pi_i.$$

(Z may be replaced by the best predictor given Z and X)

Macro editing

- Look at the final statistics. Are they reasonable? (You can e.g. find unrealistic consequences of outliers in this way or other bad estimation formulas). You know how much they usually change and in what range they should lie
- One problem is what to do if you find odd values, but does not know what it depends on. Are you allowed to change a figure that is obviously incorrect to a sensible guess?
  - The error in shoe prices in the Consumer Price Index was detected after month) (But not in a formal macro editing procedure. There was no such procedure implemented), but when they went through the raw data and the calculations they could not find the error. That took three months. And it would be a scandal if Statistics Sweden did not publish CPI for three months.
  - There was recently a similar problem with the amount of house mortgages with fixed and moving interest. Statistics Sweden suspected a reporting error from one of the banks but after a call back the bank said that their figure was correct. Are you allowed to change a reported figure just because you suspects it to be false?

4.2 Missing values - models

Daniel Thorburn

Ekonomic statistics

Autumn semester 2010

Missing values

- Unit non response (Totalbortfall)
  - i.e. no answer from the unit
- Item non response (Partiellt bortfall)
  - You have some figures for the unit, but not all.
    - The unit may not have responded to a some items in a questionnaire.
    - Sometimes it is about data collected from registers where the unit is missing in one of them or you have not asked the unit that question by design
- Strukturally missing values
    - Variables that logically cannoot exist.
    - E.g. if an enterprse has now buildings of its own iit can have now building area or no heating system
- Dark numbers
  - You do not even no that the unit exist
    - This is called frame error if the register is used for sampling. But when statistic Seden reports the number of firms or cases of bribery, or accidents related to work.

Rubin's terms

Se Little RJA and Rubin DB (2002) *Statistical Analysis with Missing Data*, 2nd edition. New York : John Wiley.

- MCAR, Missing Completely at Random
  - A value or unit is missing completely at random or another cause which is untrelated to the value
  - Formally R and Y are independent if R is the non response mechanism and Y the study variable
- If this holds you can forget about the non response

Rubin's terms

Se Little RJA and Rubin DB (2002) *Statistical Analysis with Missing Data*, 2nd edition. New York : John Wiley.

- MAR, Missing at Random
  - That a value or unit is missing depends only on circumstances ant data that we know of. The fact that it is missing does not tell us anything except what we already know
  - Formally B and Y are independent given X where X are known auxiliary variables
  - The non response may be different in different industries and number of employees but those are already known in the register and preliminary tax withdrawn from the employees and paid to the tax authorities
- Much of what is written on non-response correction relies on this. But there is always a problem to know how to use what you know. Sometimes it is enough to poststratify after some variables like industry and number of employees.

Rubin's terms

Se Little RJA and Rubin DB (2002) *Statistical Analysis with Missing Data*, 2nd edition. New York : John Wiley.

- NMAR, Not Missing at Random,
  - The non response depends on the study variables.
  - For example. It is thought
    - that expanding one-man-enterprises respond less since the entrepreneur has less time to answer
    - that firms which have large economic problems answer less since they seldom have their accounts in order.

Different ways to react when non-response occurs

- Avoid – Good questions, motivate … Draw conclusions for future data collections
- Declare – Describe the non-response, analyse, discuss possible consequencies, make follow-up studies of the non response, Quality declarations
- Try to correct (OK) and forget (not OK)
  - Reweighting (Give enterprises which have properties like many non responding units higher weight. Let it represent more firms)
  - Imputation (substitute with sensible values)

5.3 Handling data with missing values - imputation

Daniel Thorburn

Ekonomic statistics

Autumn semester 2010

(short version)

Imputation

- Very often values are missing for enterprises
  - Item non-response
  - Unit non response
- Imputation means that you enter a value in the missing value's place
  - Real donor E.g. the value last year or nearest neighbour find another company with similar values on all known variables
  - Model donor E.g. nr employes*mean wage cost per person in the sample stratum
- There is a lot of theory on how this should be done
- Very common in economic statistics –less common for individuals

Real or model donor – an exemple

Data

No	var	Sex	Salary
1	17	2	14180
2	16	1	-
3	14	-	27690
4	10	1	-
5	18	2	16189
6	10	2	23457

...

Real donor

No	var	Sex	Salary
1	17	2	14180
2	16	1	14180
3	14	1	27690
4	10	1	23457
5	18	2	16189
6	10	2	23457

...

”Nearest Neighbour”

Model donor

No	var	Sex	Salary
1	17	2	14180
2	16	1	20379
3	14	1,6	27690
4	10	1	20379
5	18	2	16189
6	10	2	23457

Mean value imputation

Real donor

- Hot - Cold deck?
  - Hot deck - imputation The value comes from the dataset itself (Quite common in the behavioural sciences)
  - Cold deck – imputation The value comes from another data set (e.g. last years value. Quite common with enterprise data)
- Where find the value in hot dec?
  - From the unit which is most close.
    - But what is close? How to choose distance measure?
    - Psykologist usually uses Euklidean distance after standardising variables.
    - In economical surveys it is more common to weight the important variables more heavily. You can alsou use methods from pattern recognition.
    - Neural networks have been suggested and works OK ...
  - From one of the 10 most similar (random imputation)
  - From the previus unit

Some remarks on model imputation

- Mean imputation
- Regression imputation. Construct a (logistic) regression model where the known auxiliary variables predicts the study variable
- May be risk for too small variance estimation and for systematic errors.
  - The imputed values lies on the regression line, which means that they should not contribute to he variance estimation.
  - E.g. if P(man|known values) = 0.7 you may impute the odd person being 70% man. Or if you select the most likely value all imputed person will be men (instead of 70 %)

- In enterprise data you have often access to last year’s data. These can often be used as explaining variables in the model. Very common is last year´s value corrected by inflation (often by using the increase among the responding units)

- Very common is also to impute missing values who can be computed logically. (Sometimes the sum in a summation is missing or ”Bengtssons car repairs Ltd” has not filled in its activity. But sometimes these are also checked by call backs)

- Imputation with extra random error is sometimes made to avoid bias and too small variance estimates. But then you introduce an extra random error and this should be done many tiames and the estimate should be the mean of them. ”Multiple imputation”

Documentation

Prevously (?) the imputation was made by experienced persons who chose very reasonable values. But the main problem is that no one knew what was done or if the result was perfect or just nonsense. It is thus very important to describe how the imputation was done.

5.4 Handling data with missing values - reweighting

Daniel Thorburn

Ekonomic statistics

Autumn semester 2010

(not included here)

6 Other Issues

Daniel Thorburn

Ekonomic statistics

Autumn semester 2010

Quality is important

- Statistics Sweden is working to get ISO-certified
- You have read quite a lot on quality in surveys (Samples and total) But quite another problem is quality in register. How do you measure the quality i the Enterprise reegister or Population register?

Relevance, validity, reliability

- Relevance – The information should be of importance for the intended use.
- Validity – Small error – The statistics should measure what you believe it does.
- Reliability – If the measurement is repeated one should get the same result. The random error is small. The purchasing manager index should not depend the weather or whether it is performed on a Monday or a Friday. – Small random error.

Quality in registers

- Qvality in registers is very important. In principle you hav all values but how good are they? It may be motivated to do a follow up survey
- The requirements in administrative registers or not the same as for statistics production. In an administrative register al items should be at least approsimately correct. In a statistical register the proportion with certain properties must be correct. The register with persons income based on their self assessment is fairly good for most individual but will result in a large underestimate of the summed income for the population.
- Previously statistical data bases were used only once and for only one purpose. Nowadays most uses are secondary linkages This means other requirements on the quality..
- Earlier students learned how to make descriptive statistics in order to interpret the data. Today it is much more important to learn how to construct the data bases so that they will be easy to use and interpret.

Timeliness

- Aktuality The information should be uptodate. It is important to present statistics fast. Sometimes much more important than a small random error.
- Punktuality is extremely important. If the statistics is promised to a certain date, those who act on it should be able to meet the day after. If you cannot rely on publication dates that may mean a postponement of months before it can be used.

Different publication times

- Forecasts (not the producer of statistics. But he may produce input to the forecasts)
- Flash estimates (Often based on material specially collected or fast counted)
- Preliminary statistics (based on all information up to a certain time)
- Final statistics
- Revisions (New information may come years later due to e.g. legal proceedings or wrong accruals. Or changes of definitions)

- Work with all these types!

Thank you for your attention