Statistics Netherlands

Division of Macro-economic Statistics and Dissemination (MSP)
Department of prices, short-term indicators and programming (MPP)

USER-DEFINED REGRESSION VARIABLES IN X-12-ARIMA

by M. Jansen

Statistics Netherlands performs seasonal adjustments with the programme X-12-ARIMA. This programme offers, besides adjustments for seasonal patterns, also the option of adjusting other effects like working day patterns, hold-ups due to frost, staggering holidays, national holidays and transition days. Some of these adjustments are executed with user-defined regression models. Generally, there are different manners to model the same effect. Unfortunately, modelling user-defined regression variables is hardly discussed in the manual of the X-12-ARIMA programme. This note describes some problems Statistics Netherlands encountered in the development of user-defined regression models and introduces the best way to model the user-defined regression effects in relation to X-12-ARIMA.

Some of the user-defined regression variables applied by Statistics Netherlands were originally defined with real values. For instance, holiday regression was modelled with the real number of bank holidays for each particular month. This has the negative effect that the whole series is forced upwards. Instead of taking real numbers it is better to base the regression variables on the deviation from average values. The overall effect on the series will then be zero and there is no upward forcing.

The use of real values causes problems when the yearly totals of the seasonal and working day adjusted series are forced to equal the original, unadjusted yearly totals. In that case, the adjusted data are revised downward. For many statistics Eurostat desires this 'alignment' procedure to avoid different weights in the aggregation of series of member states. With the 'real value' holiday regression variable,

Project number:
BPA number:
Date: *8 January 2004*

adjustments of the monthly figures after the year under review were rather high. Problems arise for the current year because the yearly total is not already known and year-on-year mutations will be adjusted considerably after the year has passed. These adjustments vary with the strength of the holiday effect which is expressed in the so-called t-value.

The problems we experienced brought us to the insight that all user-defined regression variables connected to the X-12-ARIMA programme should be based on the deviation of an average value. Regression variables should be made up by a set of positive and negative values that add up to zero. In this way, the level of the series will not change and differences between adjusted and original yearly totals will be minimal.

Of course, there are different ways to determine the average value. Research at Statistics Netherlands proved that different methods of calculating the averages will lead to more or less the same results. This pleads for a simple method, in which averages do not move in time and are based on an eternal average. As a result, the differences between original and adjusted yearly totals are small.

Appendix A gives an example of the calculation of a ‘real value’ regression variable and contains sections of the regression files used for the test in Appendix B. In Appendix B the ‘real value’ and the ‘average value’ user-defined regression variable are compared to each other by testing them on some series of the Netherlands production index.

Appendix A

This appendix describes how an 'average value' regression variable is composed. We use the example of the Dutch holiday regression variable. The 'average value' regression variable is calculated as the deviation of the number of bank holidays in a particular month from the eternal average of bank holidays for that month. We first calculate the eternal averages for each month.

Eternal averages Dutch bank holidays for each month

January: 5/7 (New Year's Day)

February: 0

March: 1.140.000/5.700.000 (Easter Monday)

April: 4.560.000/5.700.000 (Easter Monday) + 5/7 (Queen's Birthday 30th April)

May: 3.417.625/5.700.000 (Whit Monday) + 5.468.950/5.700.000 (Ascension Day)

June: 2.282.375/5.700.000 (Whit Monday) + 231.050/5.700.000 (Ascension Day)

July until November: 0

December: 10/7 (Christmas Day and Boxing Day)

The eternal averages for Easter Monday, Whit Monday and Ascension Day are calculated by means of the Easter cycle.

Easter cycle[1]

Month	Day	Frequency	Month	Day	Frequency
March	21	0	April	09	186200
March	22	27550	April	10	192850
March	23	54150	April	11	186200
March	24	81225	April	12	192850
March	25	110200	April	13	189525
March	26	133000	April	14	189525
March	27	165300	April	15	192850
March	28	186200	April	16	186200
March	29	192850	April	17	192850

[1] Source: http://members.lycos.nl/bouwzelf/paasdata.htm

March	30	189525	April	18	197400
March	31	189525	April	19	220400
April	01	192850	April	20	189525
April	02	186200	April	21	162450
April	03	192850	April	22	137750
April	04	186200	April	23	106400
April	05	192850	April	24	82650
April	06	189525	April	25	42000
April	07	189525	April	26	0
April	08	192850			

Explanation: the Easter cycle is based on exactly 5.7 million years. The frequencies of Easter Sunday for all possible dates are mentioned in the schedule above.

Easter Sunday falls on the first Sunday after the first full moon on or after the beginning of the spring. Whit Sunday is 7 weeks later than Easter Sunday. Ascension Day is 10 days before Whit Sunday.

Now we have calculated the eternal averages for every month of the year, we can easily derive the regression variable for a specific month by subtracting the eternal average from the real number of bank holidays in that month.

Example

Calculation of regression variables for Dutch bank holidays, 1990

Month	Number of holidays	Eternal average	Regression Variable
Jan	1	0,71429	0,28571
Feb	0	0	0,00000
Mar	0	0,20000	-0,20000
Apr	2	1,51429	0,48571
May	1	1,55905	-0,55905
Jun	1	0,44095	0,55905
Jul	0	0	0,00000
Aug	0	0	0,00000
Sep	0	0	0,00000
Oct	0	0	0,00000
Nov	0	0	0,00000
Dec	2	1,42857	0,57143

Finally, on the left side of the following page, we give a section of the regression file for Dutch bank holidays we just discussed. For comparison reasons, a section of the ‘real value’ regression variable is listed on the right side.

Average value regression file

1990	1	0.28571
1990	2	0.00000
1990	3	-0,20000
1990	4	0.48571
1990	5	-0.55905
1990	6	0.55905
1990	7	0.00000
1990	8	0.00000
1990	9	0.00000
1990	10	0.00000
1990	11	0.00000
1990	12	0.57143
1991	1	0.28571
1991	2	0.00000
1991	3	-0.20000
1991	4	0.48571
1991	5	0.44095
1991	6	-0.44095
⋮		
2007	6	-0.44095
2007	7	0.00000
2007	8	0.00000
2007	9	0.00000
2007	10	0.00000
2007	11	0.00000
2007	12	0.57143
2008	1	0.28571
2008	2	0.00000
2008	3	0.80000
2008	4	-0.51429
2008	5	0.44095
2008	6	-0.44095
2008	7	0.00000
2008	8	0.00000
2008	9	0.00000
2008	10	0.00000
2008	11	0.00000
2008	12	0.57143

Real value regression file

1990	1	1
1990	2	0
1990	3	0
1990	4	2
1990	5	1
1990	6	1
1990	7	0
1990	8	0
1990	9	0
1990	10	0
1990	11	0
1990	12	2
1991	1	1
1991	2	0
1991	3	0
1991	4	2
1991	5	2
1991	6	0
⋮		
2007	6	0
2007	7	0
2007	8	0
2007	9	0
2007	10	0
2007	11	0
2007	12	2
2008	1	1
2008	2	0
2008	3	1
2008	4	1
2008	5	2
2008	6	0
2008	7	0
2008	8	0
2008	9	0
2008	10	0
2008	11	0
2008	12	2

Appendix B

In this appendix both the 'real value' and 'average value' regression variables are tested on some important series of the Netherlands production index. The outcome of the test is given in the last four columns of table 1. B1 series are adjusted for working day patterns and D11 are the seasonal adjusted series. A1 is the original series. Column 2 shows the kind of regression variable applied and column 3 the corresponding t-value.

Table 1: annual totals 2002 according to the 'real value' and 'average value' method of modelling

Sbi code	Rgr[2]	t-value	2002	Real value		Average value	
			A1	B1	D11	B1	D11
15-37	HT	-7,87	1170,2	1186,3	1186,4	1173,5	1173,7
15+16	H	-3,36	1204,9	1214,7	1215,3	1205,8	1206,3
17-19	HT	-4,87	1077,9	1100,4	1100,6	1084,2	1084,8
21+22	HT	-6,26	1144,1	1164,3	1144,8	1147,9	1144,2
23-25	H	-4,97	1247,2	1258,8	1258,8	1250,1	1250,2
27-35	HT	-7,53	1125,4	1145,8	1145,9	1128,6	1128,6
20+26+36+37	H	-26,88	1158,4	1192,0	1191,8	1158,7	1158,5

Table 1 shows that annual totals with the 'average value' regression variable lie nearer to the original annual total than with the 'real value' regression. With 'real value' regression the relative differences of the annual totals are bigger as the corresponding t-value is higher. In this way, the series 20+26+36+37 will be 'aligned' and adjusted at the end of 2002 with a substantial 3 percent on average. With the 'average value' regression variable the adjustments will be much smaller and also the adjustments of the year-on-year mutations of 2002. The same table is made up for the current year 2003. For the year 2003 figures are available until October 2003 and the results are as follows:

[2] Applied kind of regression variable. H = holiday regression, HT = holiday/transition day regression.

Table 2: annual totals 2003 (until October) according to the 'real value' and 'average value' method of modelling

Sbi code	Rgr	t-value		Real value		Average value	
			A1	B1	D11	B1	D11
15-37	HT	-7,87	953,3	963,0	966,8	952,2	955,5
15+16	H	-3,36	975,2	980,3	993,0	974,1	985,8
17-19	HT	-4,87	862,2	876,3	873,3	864,3	858,2
21+22	HT	-6,26	920,1	931,2	932,9	918	930,4
23-25	H	-4,97	1078,0	1084,8	1078,0	1077,0	1068,8
27-35	HT	-7,53	901,9	915,48	917,72	900,9	904,5
20+26+36+37	H	-26,88	923,2	944,7	944,6	922,9	921,4

Once again the annual totals of the working day and seasonal adjusted series according to the 'average value' regression variable would lie much closer to the original annual totals than they are with the 'real value' regression variable. Adjustments at the end of the year will then be proportionally smaller.

Classification SBI codes:

15-37: Manufacturing industry

15+16: Food, beverages and tobacco industry

17-19: Textile, clothing and leather industry

21+22: Paper and printing industry

23-25: Oil, chemicals and rubber industry

27-35: Metal, electric engineering and transport equipment

20+26+36+37: Wood, construction materials, furniture and other industry