The Early Warning Analysis of Unemployment in the US

Author: Chenlu Wei

Wenjuan Hu

Supervisor: Mikael Elenius

C-level essay in Statistics, Fall 2006

Department of Economics and Society, Dalarna University

ABSTRACT

Researches on developing early warning system (EWS) models of currency crisis received a strong stimulus in recent years. They typically produce a predicated probability of crisis. This kind of concept is applied in the fields of unemployment in this paper. A binary response variable model is built to analyze the data from the US, predict the probability of the change of unemployment rate, and judge whether there would be an unemployment crisis.

KEY WORDS: early warning of unemployment analysis, EWS, crisis, binary response variable model

Catalogue

1 Introduction

Employment is a difficult problem that every country ought to face. A global employment report issued by the International Labor Organization on Jan. 22, 2004 shows that, the total unemployment and the unemployment rate worldwide is 1.86 million and 6.2% respectively in 2003, besides there are 7 to 9 million labor force in the state of underemployment. Researches on employment have become an important part of macroeconomics, and play an important role in keeping the society steady and the economy growing. Therefore, in order to make proper employment policy, encourage the confidence of consumers and investors, and predict the future tendency of economy, we should evaluate the state of employment precisely an in time.

Unemployment rate is an important indicator of estimating the state of employment. Changes in unemployment rate will affect the whole life cycle of economic expansions (Leamer, 2001). Economists use this indicator to make early warning analysis of macroeconomics. But, is there a method to make early warning of the rate of unemployment? Can we predict its trend before massive unemployment events take place, so that certain measures can be taken to avoid extreme unemployment situation and the breakout of social conflict that follows? These are the problems that are mainly concerned in this paper.

There have been many researches concerning on methods of early warning (Zhao 1999). The methods include calculating the Diffusion Index (DI), Composition Index (CI), and prosperity index, or making regression forecast, etc. There isn't a standard method in early warning of unemployment. The methods mentioned above have their strong suit respectively.

The method used in this paper is similar to regression forecast, the idea comes from EWS (early warning systems) models. The model I choose is binary response variable model. In this model, the main point is to forecast the direction of changes in unemployment in the state of probability, and then decide whether to give warning or not. Since to our knowledge this is the first attempt in this field, some problems rise along with building up the model, discussions will be made correspondingly.

1.1 Background

In the last ten years of 20^{th} century, the world's economy developing so fast, but along with the fast growing rate, there comes out more and more crisis. For instance, the Mexican crisis in 1994, and the Asian crisis in 1997, both damage the economy in many countries. So many countries pay more attention to the researches of early warning systems of crisis. Several EWS models have been put into use, such as Developing County Studies Division (DCSD) model, Kaminsky-Lizondo-Reinhart (KLR) model , GS-WATCH , EMRI , and Deutsche Bank Alarm Clock, etc.(Berg 2005)

Generally speaking, these EWS models have similar train of thoughts, which is to define the crisis and then select indicators correspondingly to build up models. But the crisis definition, variables, horizon and methods vary in different model. Take DSCD as an example, its crisis definition is weighted average of one-month changes in exchange rate and reserves more than 3 (country-specific) standard deviations above country average; the variables include current account, reserve losses, export growth, real interest rate, domestic credit growth, etc.; the horizon is 2 years. Compared with EMRI, the crisis definition is depreciation larger than 5% and at least double proceeding months, the variables include growth of credit to private sector, stock price growth, GDP growth, etc.; the horizon is 1 month. Though these models run in a rather complex way, the core method is quite basic, some of them employed limited response variable probit/logit models (Abiad 2003), calculating the probabilities of crisis or tranquil in a maximum-likelihood framework, and decide whether to give warnings.

Though these EWS models mentioned above is always attempt to predict currency crises, I think similar thought can be used in the analysis of unemployment.

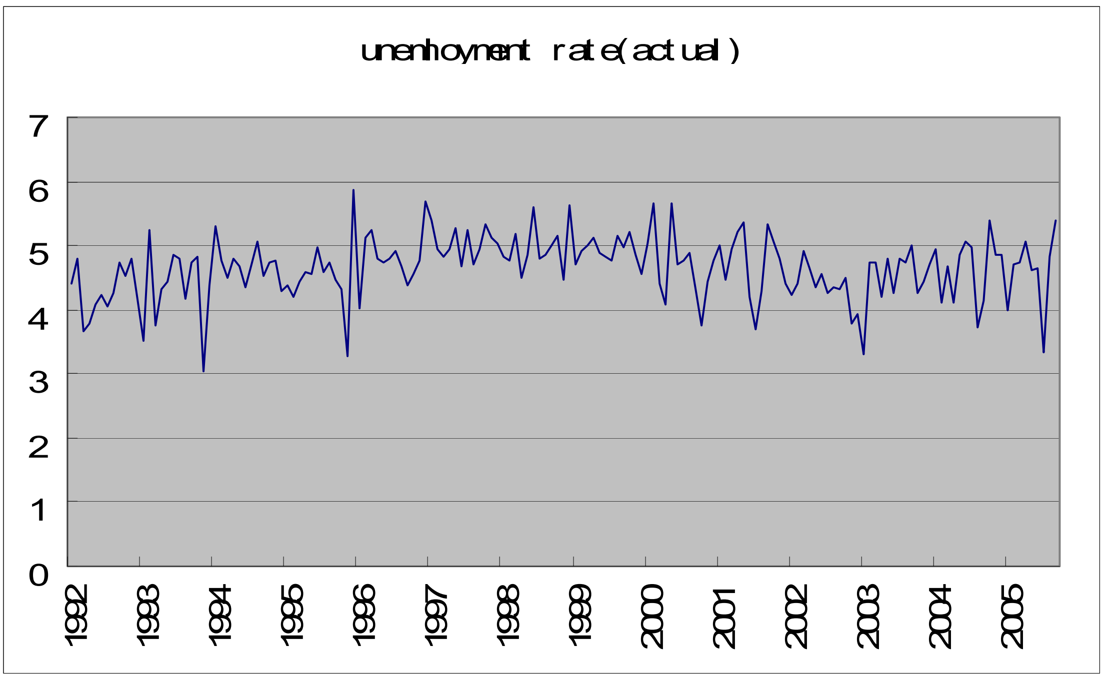
2 Data and Method

We will first try to define the unemployment crisis when the predicted probability of unemployment rate keeps on raising in some successive periods, otherwise the state of employment is tranquil. And then select a series of explanatory variables according to the significance of economic and statistic analysis, build and fit the model, for example the limited independent model, at last forecast the probability of the growth of the rate of unemployment, judge whether there is a crisis and decide whether to give out a warning.

2.1 Choose data

The rate of unemployment is a sensitive indicator. So consider the effectiveness for a given period of time, we should use monthly or seasonal data to make early warning analysis. The existing data relevant to the unemployment of the USA is easy to obtain from the official website of Bureau of Labor Statistics and US Census Bureau. And it has a rather long extent, thus they can support the analysis of early warning effectively. So I choose the monthly data of the USA to fit the model.

Figure1 Actual unemployment rate from 1992 to 2005



2.2 Build the model

As is mentioned above, our main attempt is to transform the continuing changes of unemployment into the concept of crisis. We give the definition of unemployment crisis as follows: when the rate of unemployment keeps on rising for 3 months, it is considered as a crisis. Thus we can give out a warning. In order to predict the probability of increasing in the rate of unemployment, we can employ a binary response model. In the binary response model, we model the response probabilities as functions of the predictors (Olsson 2002). Besides, we chose the Logit regression, since it is more commonly used and default in Eviews.

Let the percentage change in unemployment rate over the previous month be Y^*(column vector), then there should be a series of explanatory variables as the row vector X , they have the following relationship being the state of matrix:

$$Y^* = X\beta + \mu^* \qquad (1)$$

Besides, let Y be the observed response variable, which is determined by whether Y* exceeds a threshold value, which is 0 in the paper:

$$Y = \begin{cases} 1, Y^* > 0 \\ 0, Y^* \leq 0 \end{cases}$$

This means if unemployment rate increases compared to the previous month, the observed response variable Y would be 1; otherwise it would be 0.

Then, we have the following probability model:

$$P(Y_i = 1 | X_i, \beta) = P(Y_i > 0)$$

$$= P(\mu^* > -X_i^* \beta)$$

$$= 1 - F(-X_i^* \beta)$$

$$P(Y_i = 0 | X_i, \beta) = F(-X_i^* \beta)$$

The two probabilities are the conditional probabilities of Y being 1 and 0 respectively when the explanatory variables X being a set of values. F is the cumulative distribution function of assumed residualμ*, it is the logistic distribution. So it can also be written as:

$$P(Y_i = 1 | X_i, \beta) = 1 - e^{-x_i\beta} / \left(1 + e^{-x_i\beta}\right) = e^{x_i\beta} / \left(1 + e^{x_i\beta}\right)$$

If we set the value of X, we can work out the value of the parameter β, which is a column vector, by fitting the regression model, then forecast the state of unemployment.

To sum up, we choose the Binary Response variable Model. It is also should be aware that while we use the observed value of Y to fit the model, the parameter we work out is the value of βwhich satisfies the equation (1).

2.3 Select the explanatory variables

Some candidate indicators were selected according to certain economic and social significance. The range of these indicators is from March 1992 to May 2005. The sources of data are the Bureau of Labor Statistics and US Census Bureau. The indicators and the reason why chose them are as follows:

1. Average weekly hours of production workers: When the weekly hours keep on declining, it may predict a surge of job displacement.

2. Average weekly overtime of production workers: When the index keeps on growing for at least 3 months, enterprises would be confronted with the pressure of recruiting, which is regarded as the prelude of increasing long-term employment.

3. Index of aggregate weekly payrolls: When it grows, it means the whole industry works well, the demand gap of skilled worker is becoming larger.

4. Population of employed part time (persons who usually work less than 35 hours): When it is growing steadily, the whole economy probably is in the state of declination; when it falls, full-time job position is becoming abundance. Generally speaking, changes in part time employment may predict the future state of employment.

5. New orders for durable excluding transportation and defense: Its growth predicts a bright future of finance and employment.

6. PPI(Producer Price Index): There are PPI of crude materials and finished goods respectively. PPI (crude materials) mainly affects the monetary market, while PPI (finished goods) is very sensible to the turning point of the economy.

7. Unemployment insurance weekly claims (first claim): It is one of the premonitory indicators, as a part of the American leading indicators. In this paper, I choose the data registered in state government (SA). Since it is weekly and the series are always unsteady, we make moving average of the series(Li, 1999). See equation (2).

$$\begin{aligned}\hat{X}_t &= \frac{1}{2}\left(\frac{X_{t-2}+X_{t-1}+X_t+X_{t+1}}{4}+\frac{X_{t-1}+X_t+X_{t+1}+X_{t+2}}{4}\right)\\ &= \frac{1}{8}\left(X_{t-2}+2X_{t-1}+2X_t+2X_{t+1}+X_{t+2}\right)\end{aligned} \tag{2}$$

Thus we work out the average of every month, eliminate the fluctuation, and generate the monthly series.

The above monthly data and the unemployment rate are all standardized, using March 1992 as the benchmark. Every datum is divided by the corresponding datum in March 1992. With the standardized data and the 0/1 variable Y that estimates the changes of unemployment, we can fit the model. The sample used to build the model and make fitness test is ranged from March 1992 to May 2004; the range of out-of-sample forecast is from June 2004 to May 2005.

We put all the variables into the Binary Response variable Model, and then we revise it time by time to get rid of the insignificant variables, and do the goodness-of-fit measures. Eviews provides several statistics and methods for Goodness-of-fit Test. In this paper, we employ the value of McFadden R.squared, AIC and SIC, Hosmeer-Lemeshow test and Andrews test.

McFadden R-squared is the likelihood ratio index. As the name suggests, this is an analog to the R^2 reported in linear regression models. The values of AIC and SIC, which is an operational way of trading off the complexity of an estimated model against how well the model fits the data, are also taken into consider. The Appendix contains two output results of EVIEWS, including the final model, and one of the abandoned candidate models. We can see that the value of McFadden R.squared in the final model is 22%, while the alternative one is 23.7%. They do not differ a lot. That means the forecasting ability is also equivalent to one another. But comparing to the alternative model, the final one has lower AIC and SIC, along with much lower significant probability of every coefficient. So the final one is simpler and has clearer relationships between the response variable and explanatory variables.

Therefore, we finally establish the following model including 6 variables. It is shown in Table 1.

Table 1 Estimated coefficients of model

Variable	Coefficient	Z-Statistic	Prob.
Unemployment insurance weekly claims (first claim) (-1)[1]	29.70643	4.151560	0.0000
New orders for durable excluding transportation and defense (-12)	-16.51073	-2.920053	0.0035
Population of employed part time (-1)	-35.66935	-3.212080	0.0013
Index of aggregate weekly payrolls (-12)	5.857404	2.310129	0.0209
Unemployment rate (-2)	32.39586	2.323460	0.0202
Unemployment rate (-1)	-69.99301	-4.014817	0.0001

Table 2 is the output of Hosmer-Lemeshow test and Andrews test. The null hypothesis is that the fit is sufficient to the data. With the p-value (shown in the left part of the table), we cannot reject the null hypothesis, so it is reasonable to consider the

[1] The numbers in the bracket represent the lagging periods. In other word, notate $X(t)$ as a time series of data, set a stage you are now looking at, then $X(-1)$ stands for the preceding stage, and –1 means it is one step before the present stage, etc.

goodness-of-fit is quite acceptable, and expect a well-performed forecast ability.

Table 2 The output of goodness-of-fit

	Chi-Sq Statistic	df	P-value
H-L Statistic	3.1674	8	0.9234
Andrews Statistic	13.6922	10	0.1875

3 Result

3.1 The explanation of the model

Here are the explanations to the model:

1. Estimated coefficients from a binary model cannot be interpreted as the marginal effect on the response variable. The direction of the effect of a change in an explanatory variable depends only on the sign of its coefficient. Positive values of the coefficient imply that increasing explanatory variable will increase the probability of the response; negative values imply the opposite.

2. The change in unemployment has the character of inertia. The change in current period is relative to the one in previous periods. According to many researches, in the recovery period of the economy, if the rate of unemployment declines, it would decline further in the several periods. Whereas during the spurt of growth, the unemployment rate always shows opposite state in the previous and lag period. For instance, we can find proof in this model, when the other explanatory variables keep constant; the direction of change in unemployment rate in current period is opposite to the previous one period, and the same as the lagged two periods. We also use PACF and ACF to exam the unemployment rate, which shows the character of AR(2). This situation matches the state of American economy in the end of 90's of last century, because the policies executed by the Clinton government and Greenspan lead the American economy to a fast growing period. So the state of unemployment that shows

in the sample is quite normal.

3. Further explanation to other indicators. In traditional dynamic analysis of the labor market, the main factor is productivity. Besides, other factors include the cost of employment, unemployment insurance, etc. So the economic significance of the indictors in the model is based on these factors. The indicator that represents the state of productivity include new orders for durable excluding transportation and defense; the one that represent the employment cost include index of aggregate weekly payrolls; the one that stands for unemployment insurance is unemployment insurance weekly claims (first claim). Besides the traditional dynamic analysis of the labor market, there are other indictors, which cannot explain or affect the state of unemployment directly, but closely related to the state, can reflect the change in demand and supply of labor sensitively, for example, the population of employed part time. The indicators mentioned above combine together, thus explain the change in unemployment rate.

3.2 The forecasts

With the definition given above, we can now use the model to forecast. The probability we work out is:

$$\hat{p} = 1 - F(-x_i^* \hat{\beta})$$

The series YF is the conditional probability of the event that the unemployment rate would go up. In Chart 1, the unshaded area is the value of prediction (March 1992 to May 2004), and the shaded area is the value of forecast (June 2004 to May 2005).

Figure 1　　Forecasts of probability

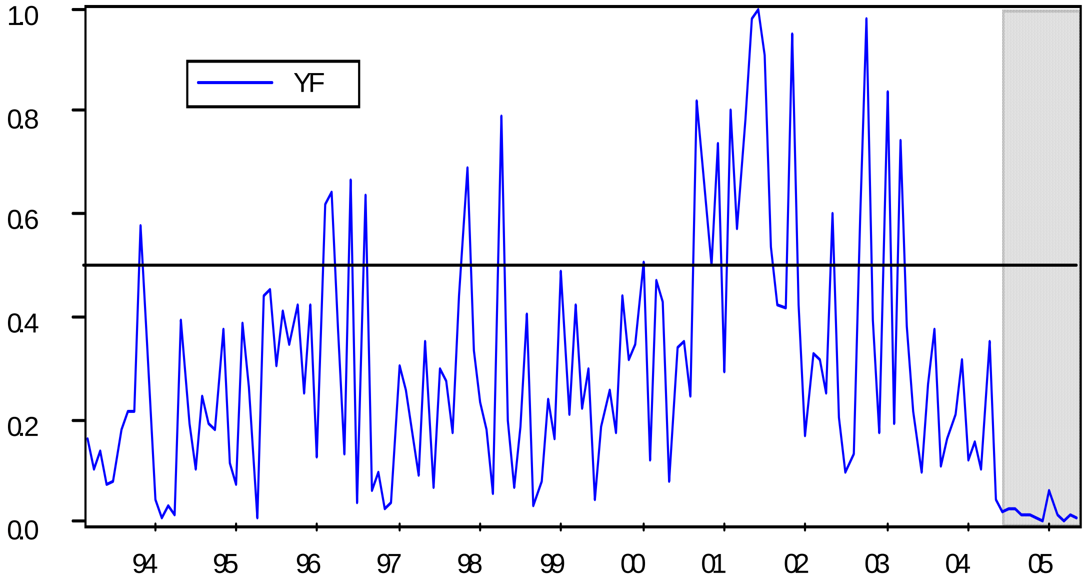
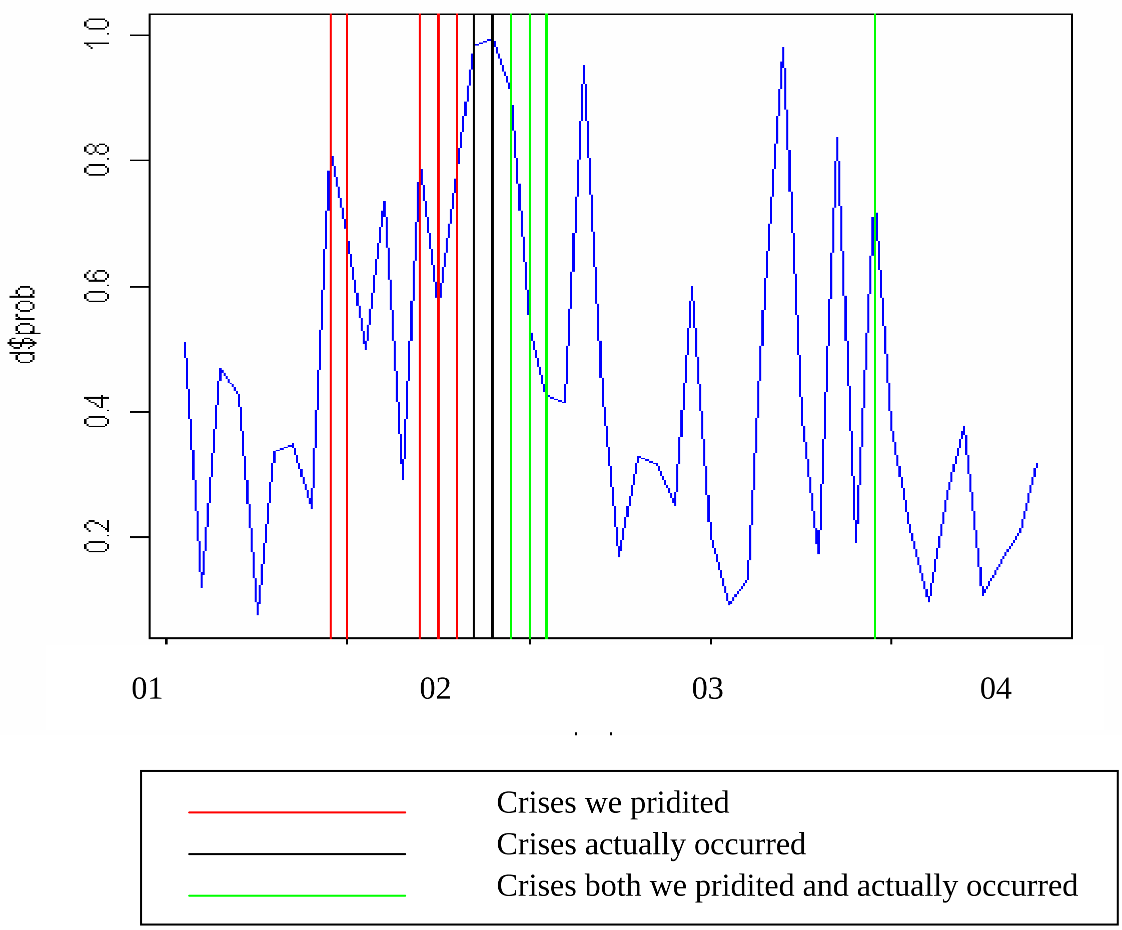


Figure 2　　Forecasts of probability(show the positions of crises)



We can assume when the conditional probability of the event that the unemployment rate goes up is over 50% in three successive months, the predict unemployment rate goes up in three successive months, too. Then the alarm signal would be sent out. The black line in the above chart is the threshold of alarm. Now we will compare the in-sample and out-of-sample early warning with the actual unemployment crisis.

According to the definition we assumed at the beginning, when the rate of unemployment keeps on rising for 3 months, it is considered as a crisis, thus we can give out a warning. There are six actual crises from March 1993 to May 2004. We get seven warnings from in-sample forecast. And there are neither actual crises nor warnings in out-of-sample forecast, we can expect a good status of the labor market in the US. Table 3 is the result of comparison.

In November and December of 2000, the model gives out early warning, though actually there is no defined crisis happen, the unemployment rate began to rise from December 2000, and continued rising in two successive periods, February and March of 2001. The last five warnings are 3 months prior to the actual crises: the predicted period is from April to August 2001, while the actual crisis is from July to November 2001. At last, there is a defined crisis actually happened in May 2003, meanwhile the predicted probability reached 83.6% and 74% in January and March of 2003, although it did not show as a crisis.

Since there aren't too many actual crises and warnings, it is not quite reasonable to judge the efficiency of prediction through the percent of crises correctly called. But it can be observed from the data that the model estimates the trend of unemployment quite well, and reach good effect of predict mass crises.

On the other hand, EVIEWS provides Expectation-Prediction table to analyze the effect of forecast. Table 4 shows the output. Here we set the cutoff be 0.5, equal to the threshold of the model. In the upper left-hand table, observations are classified as having predicted probabilities that are above or below the cutoff value. In the upper right-hand table, we classify observations using the predicted probability given by the sample proportion of y=1 observations. In the upper part of the table, the Total Gain shows that the estimated model correctly predicts have 7.41% increases, the Percent

Gain is 23.26%.

In the lower part of the table there are comparison of the expected number of y=0 and y=1 observations. The improvement is 10.97 percentages point over the constant probability model; the Percent Gain is 25.26%. So generally speaking, current model have quite good effect of forecast.

Table 3　The output of E-P table when the cutoff is 0.5

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	85	26	111	92	43	135
P(Dep=1)>C	7	17	24	0	0	0
Total	92	43	135	92	43	135
Correct	85	17	102	92	0	92
% Correct	92.39	39.53	75.56	100.00	0.00	68.15
% Incorrect	7.61	60.47	24.44	0.00	100.00	31.85
Total Gain*	-7.61	39.53	7.41			
Percent Gain**	NA	39.53	23.26			

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
E(# of Dep=0)	70.10	21.90	92.00	62.70	29.30	92.00
E(# of Dep=1)	21.90	21.10	43.00	29.30	13.70	43.00
Total	92.00	43.00	135.00	92.00	43.00	135.00
Correct	70.10	21.10	91.20	62.70	13.70	76.39
% Correct	76.19	49.07	67.55	68.15	31.85	56.59
% Incorrect	23.81	50.93	32.45	31.85	68.15	43.41
Total Gain*	8.05	17.21	10.97			
Percent Gain**	25.26	25.26	25.26			

Summary

We try to use the probability to estimate the changes of unemployment, so as to use the probability to estimate the crisis. The model is not perfect, but it shows the above thought quite well. So the model can be a good attempt of early warning analysis.

Within the whole process of analysis, there are several problems needed further discussion.

1. Problems of the definition of crisis and the threshold.

In this model, the defined crisis happens when the unemployment rate keeps on rising for 3 months; the early warning was sent out when the conditional probability of unemployment rate grows over 50% for 3 successive months. But we did not make full discussion about the best definition the threshold.

Those in different EWS models do not have standard rules. Take the threshold for example, it is mentioned in Abdul Abiad (2003), different economists pick different threshold for analysis. Different threshold will lead to different early warnings.

So it is not a problem that have been settled, we ought to decide according to the actual situation and the feature of the model.

2. Problem about the auto-correlation in the binary choice model.

As we have discussed above, unemployment has the character of inertia and AR(2). So we employ Y(t-1) and Y(t-2) in the model. In other words, large proportion of the prediction is dependent on the lag values of Y. Is there a way to improve the current model without considering the auto-correlation of response variable? We would try to find more efficient explanatory variables in future research.

3. Problem about the in-sample forecast and out-of-sample forecast.

The effect of out-of-sample forecast is one of the best methods to evaluate the model. But the length of data in this model is limited. In order to promise the effect of goodness of fit in the sample, the length of out-of-sample forecast was reduced. Though it is shown above that the probability of crisis in short term is quite low, it is a pity we cannot make long-term prediction.

Out-of-sample forecast is an important problem in the research of EWS models. Berg (2005) pointed out, even in standard model like DCSD, the percentage of false alarm is higher out-of-sample than in-sample. Therefore, we should make further discussion of increasing the quality of out-of-sample forecast, and attempt to employ longer range of data to the model to get further and better forecast in future research.

4. Problems of the forecast of rare disaster

Rare disaster is the unusual or extreme event, for instance the war, nature calamity, and terrorist attack, etc. The rare disasters do not happen frequently. Some economists are now trying to explain how they affect the earning rate of stocks and bonds in the past 100 years. They also use probability to estimate the event. The rough method is to use the rare disaster as an explanatory variable in the model. When a rare disaster happens, the probability of future disaster becomes larger, thus affect the whole model. Similarly, in this paper, the "9.11" happened in the sample, so the events like terrorist attack have been taken in consideration in the estimated probability of unemployment rate increase after Sep.2001.

But the question is, if the events like "9.11" did not happen in the sample, how can we take them in consideration in the model? The probability function of the history data which is in a tranquil period cannot fully reveal the risk of disasters. This is a problem that cannot be avoided by using the history data. We need to find an effective method to solve the problem.

REFERENCES

1. Edward E. Leamer, *"The Life Cycle of US Economic Expansions"*, [http://www.nber.org/papers/w8192]
2. YanYun Zhao, (1999) , *"Statistics Analysis of Macroeconomics"*, China Renmin University Press
3. Andrew Berg, Eduardo Borensztein,Catherine Pattillo, (2005), *"Assessing Early Warning Systems: How Have They Worked in Practice?"*, IMF Staff Papers, Vol.52, No.3
4. Abdul Abiad, (2003), *"Early-Warning Systems: A Survey and a Regime-Switching Approach"*,IMF Working Paper
5. Ulf Olsson,(2002) *"Generalized Linear Models and Applied Approach"*, Studentlitteratur, Lund
6. Edward E.Leamer, (2001) *"The Life Cycle of US Economic Expansions"*, National Bureau of Economic Research, Mar.
7. Bernard Baumohl, (2005), *"The Secret of Economic Indicators"*, China Renmin University Press,
8. DanHui Yi, (2002) *"Data Analysis and Application of Eviews"*, China Renmin University Press,
9. Li XinYu, (1999) *"Applied Economic Statistics"*, Peking University Press,

APPENDIX

1. The Final Model

Dependent Variable: Y

Method: ML - Binary Logit

Date: 04/18/06 Time: 15:02

Sample(adjusted): 1993:03 2004:05

Included observations: 135 after adjusting endpoints

Convergence achieved after 7 iterations

Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	56.66379	16.89764	3.353355	0.0008
STCLAIM(-1)	29.70643	7.155487	4.151560	0.0000
STDURA(-12)	-16.51073	5.654255	-2.920053	0.0035
STPAYROLL(-12)	5.857404	2.535531	2.310129	0.0209
STPART(-1)	-35.66935	11.10475	-3.212080	0.0013
STUNEMSE(-2)	32.39586	13.94293	2.323460	0.0202
STUNEMSE(-1)	-69.99301	17.43367	-4.014817	0.0001

Mean dependent var	0.318519	S.D. dependent var	0.467637
S.E. of regression	0.415133	Akaike info criterion	1.079043
Sum squared resid	22.05890	Schwarz criterion	1.229687
Log likelihood	-65.83539	Hannan-Quinn criter.	1.140260
Restr. log likelihood	-84.47594	Avg. log likelihood	-0.487670
LR statistic (6 df)	37.28110	McFadden R-squared	0.220661
Probability(LR stat)	1.55E-06		
Obs with Dep=0	92	Total obs	135
Obs with Dep=1	43		

2. The Alternative Model (Used for comparison)

Dependent Variable: Y

Method: ML - Binary Logit

Date: 12/04/06 Time: 00:39

Sample: 1993:03 2004:05

Included observations: 135

Convergence achieved after 8 iterations

Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	24.72347	52.38393	0.471967	0.6370
STCLAIM(-1)	32.28197	8.574290	3.764973	0.0002
STDURA(-12)	-18.67122	6.136591	-3.042605	0.0023
STHOURS(-1)	39.94121	54.45136	0.733521	0.4632
STOVERTIME(-1)	-4.136790	8.655286	-0.477950	0.6327
STPART(-1)	-37.72197	13.47215	-2.799996	0.0051
STPAYROLL(-12)	7.226691	3.124527	2.312891	0.0207
STPPIC(-1)	0.120308	0.094884	1.267951	0.2048
STPPIF(-1)	0.013726	0.055700	0.246429	0.8054
STUNEMSE(-2)	34.16962	14.44411	2.365644	0.0180
STUNEMSE(-1)	-75.00054	18.39263	-4.077750	0.0000

Mean dependent var	0.318519	S.D. dependent var	0.467637
S.E. of regression	0.415426	Akaike info criterion	1.117772
Sum squared resid	21.39981	Schwarz criterion	1.354498
Log likelihood	-64.44959	Hannan-Quinn criter.	1.213971
Restr. log likelihood	-84.47594	Avg. log likelihood	-0.477404
LR statistic (10 df)	40.05270	McFadden R-squared	0.237066
Probability(LR stat)	1.66E-05		
Obs with Dep=0	92	Total obs	135
Obs with Dep=1	43		