Alexander Todorov
Regression Analysis VEE, Fall 2008
Final Project
Feb 18, 2009

Introduction

In this project, I will use incurred Property Damage (PD) claim frequency data and fit two types of models: exponential fit over time and regression model using two independent variables. I will then focus on the residuals from the regression model and use moving average (MA) correction for serial correlation to improve the Durbin-Watson statistic. Finally, I will generate an ex-ante conditional forecast and compare the trend obtained this way to the constant growth rate of the exponential model.

All work is done with Excel and EViews.

Models

Exponential Model:

Traditional actuarial methods use exponential fits to predict claim frequency and severity trends. The insurance data (Y) is assumed to be an exponential function of time (t):

$$Y = f(t) = \alpha e^{\beta t}$$

The parameters α and β maximize the correlation between f(t) and Y, and are estimated by taking logarithms of both sides and fitting the log-linear regression equation:

$$lnY = \ln\left(\alpha e^{\beta t}\right)$$

$$lnY = \ln \alpha + \beta t$$

or

$$Y' = \alpha' + \beta t$$

where $\ln Y = Y'$, $\alpha' = \ln\alpha$, and t is time. The estimates of α and β are obtained by the Ordinary Least Squares procedure applied to lnY. The exponential method can be highly accurate but it depends on the assumption that future trends will resemble those from the past. It also does not take into account changing economic circumstances which may have significant effect on the insurance data.

In general, claim frequency data for Bodily Injury and Property Damage is known to be quite variable and as a result the exponential curve does not fit the data well over a long period of time.

Regression Model:

The regression model that I will present, assumes that frequency data is dependent on economic conditions. In particular, I will be exploring the effect of unemployment and relative price of gasoline on the claim frequency.

Model 1:

In this model, I will use both independent variables in the regression:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Model 2:

In this model, I will combine both independent variables into a weighted average variable Z and use:

$$Y_i = \alpha + \beta Z_i + \varepsilon_i$$

Here $Z = w_1 X_1 + w_2 X_2$ and $w_1 + w_2 = 1$. The weights, w_i, are determined using Solver such that the correlation between Z and Y is maximized.

In order to choose between *Model 1* and *Model 2*, I will examine their statistical properties.

Analysis

Exponential Model:

The incurred PD frequency data (INC_PD_FREQ) covers the time period from the fourth quarter of 2001 to the third quarter of 2007 – a total of 24 points. Below are the results from the exponential fit which I did with Excel. The variable TIME runs from 1 to 24.

Regression Statistics	
Multiple R	0.9774
R Square	0.9552
Adjusted R Square	0.9532
Standard Error	0.0158
Observations	24

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.1165	0.1165	469.4554	0.0000
Residual	22	0.0055	0.0002		
Total	23	0.1220			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.4709	0.0066	221.5670	0.0000	1.4571	1.4846	1.4571	1.4846
TIME	-0.0101	0.0005	-21.6669	0.0000	-0.0110	-0.0091	-0.0110	-0.0091

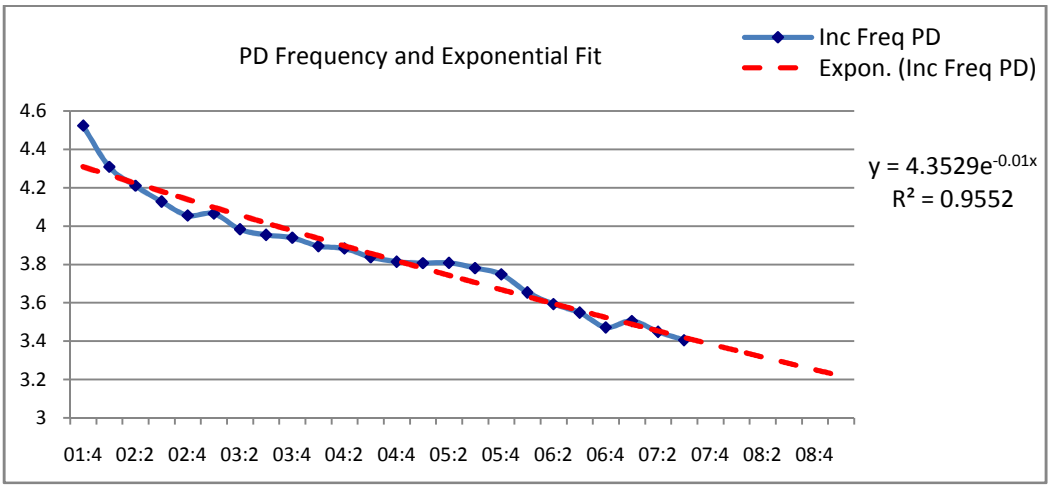
The fitted model is:

$$ln(\text{INC\_PD\_FREQ}_t) = 1.4709 - 0.01t$$

And equivalently, in exponential form:

$$\text{INC\_PD\_FREQ}_t = 4.3529e^{-0.01t}$$

Thus, the claim frequency data grows at a constant rate of -1.00%. (This is a decline of 1.00% in every quarter). The following graph shows the actual data and the exponential fit.



Regression Model:

The regression models for PD claim frequency includes four quarter moving average of the civilian unemployment rate (RUC_MA) and a four quarter moving average measure of the relative price of gasoline (RELGAS_MA), defined as the ratio of the price deflator for consumption of gasoline and oil to the price deflator for personal consumption expenditures, as independent variables.

Both the unemployment rate and the relative price of gasoline are expected to be inversely related to the incurred PD frequency data. If there are fewer people traveling to work, traffic density is reduced, lowering the likelihood of accidents. When the unemployment rate is high, the volume of business activity

is low, which results in a decrease in the usage of commercial vehicles and, therefore, we can expect lower claim frequency. When gasoline prices are high compared to prices in general, people and businesses will both seek to reduce their driving, reducing traffic density, and claim frequency. When gasoline prices are low, the amount of driving and the likelihood of accidents may increase.

Keeping that in mind, in a model such as

$$log(\text{INC\_PD\_FREQ}_i) = \alpha + \beta_1 log\ (\text{RUC\_MA}_i) + \beta_2 log(\text{RELGAS\_MA}_i) + \varepsilon_i$$

we would expect negative coefficients β_1 and β_2. If that is not the case, even if the model is statistically sound, we should discard it as there will undoubtedly be some shortcoming in the theory, data, specification, or estimation procedure.

So here is the estimation done with EViews:

Dependent Variable: LOG(INC_PD_FREQ)
Method: Least Squares
Sample (adjusted): 2001Q4 2007Q3
Included observations: 24 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.8152	0.146938	12.35349	0.0000
LOG(RELGAS_MA)	-0.4103	0.040943	-10.02145	0.0000
LOG(RUC_MA)	-0.2504	0.08582	-2.917872	0.0082

R-squared	0.8802	Mean dependent var	1.3450
Adjusted R-squared	0.8687	S.D. dependent var	0.0728
S.E. of regression	0.0264	Akaike info criterion	-4.3155
Sum squared resid	0.0146	Schwarz criterion	-4.1682
Log likelihood	54.7861	Hannan-Quinn criter.	-4.2764
F-statistic	77.1093	Durbin-Watson stat	0.5289
Prob(F-statistic)	0		

We can see from the table above is that both variables have the correct sign in the coefficients; both are significant at the 99% level, (the Prob. value is less than 0.05), and the R^2 is 0.88. However, what is not that great here is the Durbin-Watson (DW) statistic of 0.53. A DW statistic of less than 2 indicates positive serial correlation in the residuals (DW statistic greater than two indicates negative serial correlation).

Multicollinearity is something that we should always keep in mind when working with more than one independent variable. Typically, high standard errors with low t-statistics could be indicative of multicollinearity. Multicollinearity would also occur if the explanatory variables are sufficiently highly correlated. This would make it difficult to separate the effects of one explanatory variable on the dependent variable from the effects of the other explanatory variables.

If we focus on the standard errors and t-statistics of this estimation, we would not suspect multicollinearity, but a quick check of the correlation between RELGAS_MA and RUC_MA reveals that they are relatively highly correlated[1]:

$$\rho = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2(Y - \bar{Y})^2}} = -0.80$$

Here, X and Y are respectively the two independent variables in the regression model.

So, keeping that in mind, I introduced an MA term in attempt to improve the serial correlation in the residuals. This time, the estimation produced a model that is not statistically sound:

Dependent Variable: LOG(INC_PD_FREQ)
Method: Least Squares
Sample (adjusted): 2001Q4 2007Q3
Included observations: 24 after adjustments
Convergence achieved after 11 iterations
MA Backcast: 2001Q3

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.544504	0.186609	8.276679	0.0000
LOG(RELGAS_MA)	-0.34061	0.051496	-6.61426	0.0000
LOG(RUC_MA)	**-0.09506**	**0.108569**	**-0.87553**	**0.3917**
MA(1)	0.965875	0.02451	39.40711	0.0000

R-squared	0.961382	Mean dependent var	1.345022
Adjusted R-squared	0.95559	S.D. dependent var	0.072828
S.E. of regression	0.015348	Akaike info criterion	-5.36471
Sum squared resid	0.004711	Schwarz criterion	-5.16836
Log likelihood	68.37647	Hannan-Quinn criter.	-5.31262
F-statistic	**165.9659**	Durbin-Watson stat	1.574768
Prob(F-statistic)	0		
Inverted MA Roots	-0.97		

Here, the LOG (RUC_MA) term has become insignificant.

The *t*- statistic is defined as:

$$t_{N-3} = \frac{\hat{\beta} - \beta_0}{S_\beta}$$

[1] The formula used is Excel's CORREL(X,Y).

It is used to test the null hypothesis that $\beta = 0$. This estimation essentially tells us that we cannot reject the null hypothesis, or equivalently, that there is no relationship between the unemployment variable and the PD claim frequency. This is somewhat counterintuitive considering the previous estimation with no MA term.

One other thing that I noticed from this estimation is that the F-statistic has increased from 77.1 to 166. While higher F statistic is what we would hope to obtain, when it is in combination with low t-value, it could be an indicator of multicollinearity.

Having confirmed my suspicions about multicollinearity, I decided to introduce a new variable as a weighted average of the two variables used in the model:

$$RELGAS\_RUC = w_1\, RELGAS\_MA + w_2\, RUC\_MA$$

I wanted to maximize the correlation between RELGAS_RUC and the PD claim frequency data, so I used Excel's Solver subject to the following constraints:

$$0 \leq w_i \leq 1$$

and

$$w_1 + w_2 = 1$$

Solver came out with the following values: $w_1 = 0.84$ and $w_2 = 0.16$. The correlation turned out to be -0.95.

Next, I proceeded with estimating the model:

$$log(\text{INC\_PD\_FREQ}_i) = \alpha + \beta log\ (RELGAS\_\text{RUC}_i) + \varepsilon_i$$

Dependent Variable: LOG(INC_PD_FREQ)
Method: Least Squares
Sample (adjusted): 2001Q4 2007Q3
Included observations: 24 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.817543	0.032514	55.90083	0.0000
LOG(RELGAS_RUC)	-0.80951	0.055135	-14.6823	0.0000

R-squared	0.907396	Mean dependent var	1.3450
Adjusted R-squared	0.903187	S.D. dependent var	0.0728
S.E. of regression	0.02266	Akaike info criterion	-4.6567
Sum squared resid	0.011297	Schwarz criterion	-4.5586

Log likelihood	57.88098	Hannan-Quinn criter.	-4.6307
F-statistic	215.5703	Durbin-Watson stat	0.5516
Prob(F-statistic)	0		

In this model, the new variable RELGAS_RUC takes the appropriate sign, $R^2 = 0.91$, but the DW statistic of 0.55 is still too low. So again, focusing on improving the serial correlation in the residuals, I introduced an MA term in the model:

Dependent Variable: LOG(INC_PD_FREQ)
Method: Least Squares
Sample (adjusted): 2001Q4 2007Q3
Included observations: 24 after adjustments
Convergence achieved after 17 iterations
MA Backcast: 2001Q3

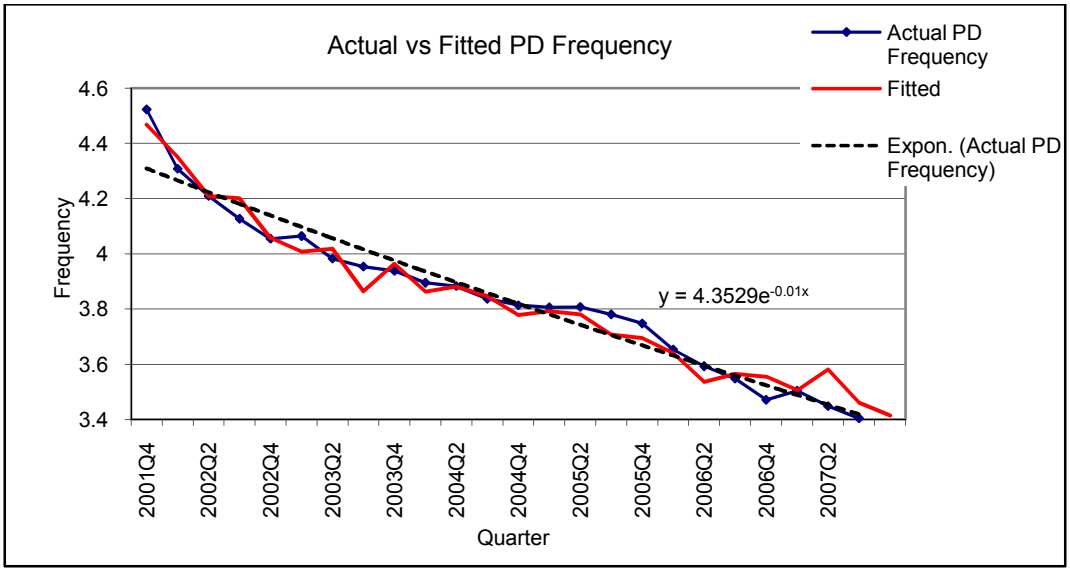
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.791431	0.043699	40.99484	0.0000
LOG(RELGAS_RUC)	**-0.768489**	0.073427	-10.46609	0.0000
MA(1)	0.948525	0.035537	26.69143	0.0000

R-squared	**0.960762**	Mean dependent var	1.3450
Adjusted R-squared	0.957025	S.D. dependent var	0.0728
S.E. of regression	0.015097	Akaike info criterion	-5.4321
Sum squared resid	0.004787	Schwarz criterion	-5.2849
Log likelihood	68.18533	Hannan-Quinn criter.	-5.3930
F-statistic	257.0998	Durbin-Watson stat	**1.65794**
Prob(F-statistic)	0		
Inverted MA Roots	-0.95		

Now, this model appears to be statistically sound. The R^2 has improved to 0.96 and the DW statistic has improved to 1.66, narrowing the gap between the statistic's value and 2, the indication of no serial correlation. The specification is as follows:

$$log(\text{INC\_PD\_FREQ}_i) = 1.79 + 0.77log\ (RELGAS\_\text{RUC}_i) + 0.95\varepsilon_{i-1} + \varepsilon_i$$

Here's how this model looks compared to the actual data and the exponential fit:



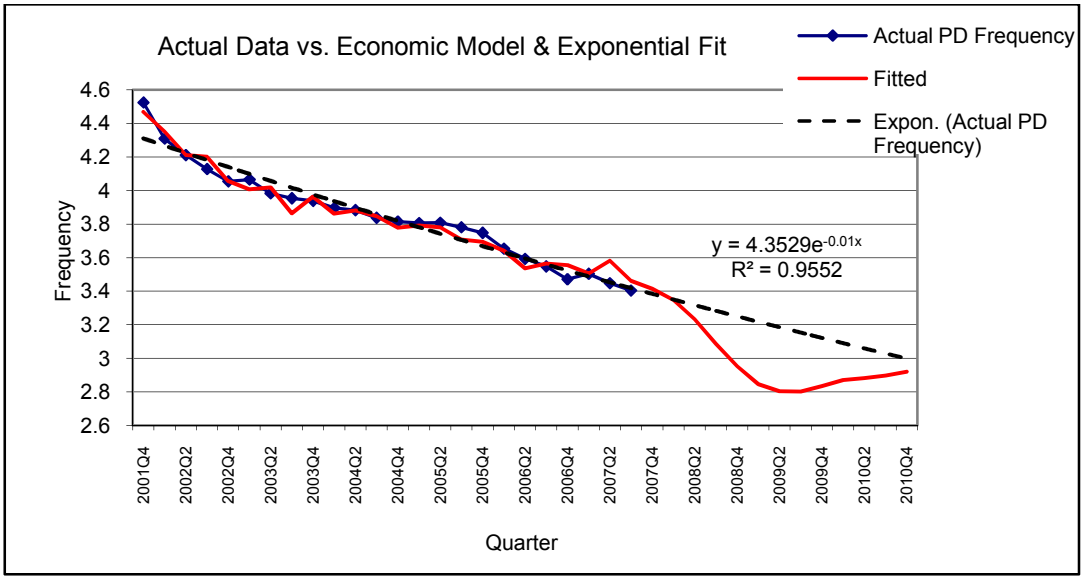
The econometric model attempts to follow the shape of the actual frequency curve, but we can see that the exponential fit is also pretty good.

Forecast

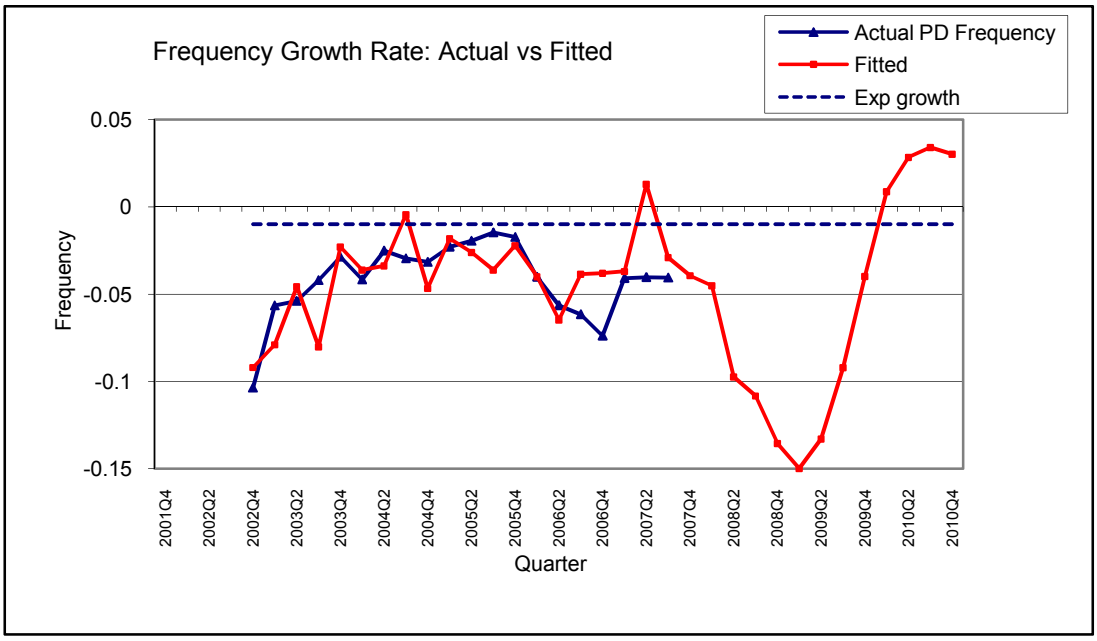
In order to forecast the model

$$log(\text{INC\_PD\_FREQ}_i) = 1.79 + 0.77log\ (RELGAS\_\text{RUC}_i) + 0.95\varepsilon_{i-1} + \varepsilon_i$$

we need to have available forecasts for the components of the composite variable RELGAS_RUC. The forecast then would be ex-ante conditional – it will depend on the forecast of RELGAS_RUC. I was able to obtain forecasts for the two components of this variable, so forecasting the PD claim frequency was straightforward. The following graph illustrates the forecast:



We can see considerable difference between the exponential line and the econometric model. The model appears to predict a greater decline in the claim frequency till about the mid of 2009, after which there seems to be an increase. Examining further the % change of the actual frequency data versus the % change of the fitted values, we see sharp decline followed by sharp increase. On the other hand, the exponential trend is constant:



For a typical actuarial scenario assuming “trend from” and “trend to” of fourth quarter of 2006 and second quarter of 2010, respectively, the model predicts a 5.8% decline in the frequency. If we use the fourth quarter of 2007 as the selected “trend from”, the model predicts a decline of 6.6%. Again, these numbers are in sharp contrast to the 1.0% decline according to the exponential fit.