

Problem Set 1 – Suggested Answers

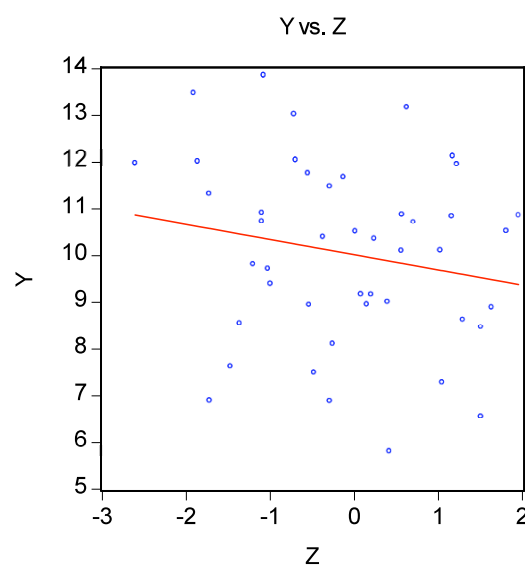
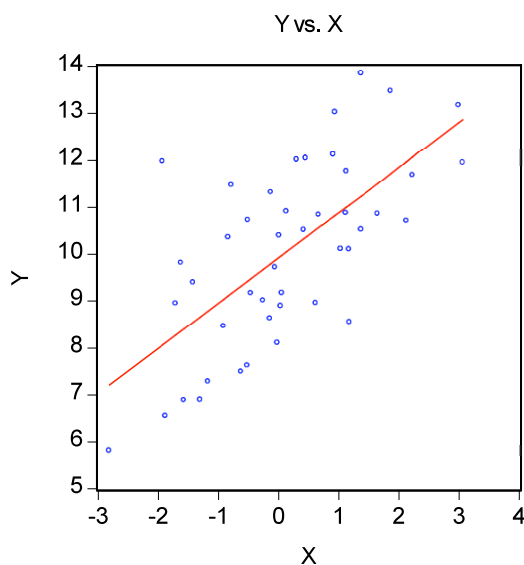
Ch 1. Exercises, Problems and Complements:

2. (p. 9) What sorts of forecasts would be useful in the following decision-making situations? Why? What sorts of data might you need to produce such forecasts?

- (a) Shop-all-the-time Network needs to schedule operators... In the past, students have mentioned a number of things one would want to forecast, including the deterministic daily pattern of high and low activity, predictable seasonal patterns in demand, developments in income, relative prices and other macro variables that might drive consumer demand for the product, etc. Data on the same items historically would be needed to estimate forecasting models for these things.
- (b) U.S. investor holding Japanese, British, French and German stocks and considering adding Tambian stocks... In the past, students have mentioned the desirability of trying to estimate an event model that would help to estimate the probability both of a coup and of devaluation. They also noted the difficulty in doing so, since there may not have been any prior experience with such events in this country; some suggested looking at experiences in other countries to try to estimate these probabilities. A model of devaluation probabilities would presumably be driven in part by macro variables in the country, so data on these would be needed. Others noted that one would want to estimate the expected return of the stocks in question and perhaps their correlation with the existing portfolio since that will determine their usefulness in your portfolio.

Ch 2. Exercises, Problems and Complements:

7. (p. 31) Using the dataset "...\\Data Sets\\fcst4-02\\PC2_7_xyz.dat" repeat the analysis of regression results from the book.



y is positively correlated with x and negatively correlated with z.

Dependent Variable: Y
 Method: Least Squares
 Date: 09/08/07 Time: 16:14
 Sample: 1960 2004
 Included observations: 45

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.848730	0.197733	49.80815	0.0000
X	1.084771	0.152900	7.094655	0.0000
Z	-0.618136	0.178510	-3.462761	0.0012
R-squared	0.561873	Mean dependent var		10.04945
Adjusted R-squared	0.541010	S.D. dependent var		1.936905
S.E. of regression	1.312230	Akaike info criterion		3.445674
Sum squared resid	72.32180	Schwarz criterion		3.566118
Log likelihood	-74.52765	F-statistic		26.93132
Durbin-Watson stat	1.469258	Prob(F-statistic)		0.000000

The table indicates that the variable y was regressed on a constant and the two “independent” variables x and z. The regression used 45 annual observations on each variable, from 1960 through 2004 (you may have set up a different date range). The “after adjusting endpoints” notation indicates that eviews may have adjusted the sample period to match the available data period.

The coefficients on the constant, x and z are all significantly different from zero at the 95% confidence level. We know that since the T-statistics are much larger than 2 in absolute value. There are two other ways to see this. First, coef +/- 2*standard error gives approximately a 95% confidence interval around the coefficient point estimates, so for example there is a 95% chance that the true value of the z coefficient is between $[-.618-2*.179, -.618+2*.179] = [-0.976, -0.260]$, which does not include zero. Finally, the Prob. column gives the probability that the true coefficient value is zero given the regression estimate, which in each case is very, very small.

The coefficient values indicate that a one unit increase in X, holding fixed Z, would lead to about a one unit increase in y. (In fact, we cannot reject the hypothesis that the coefficient is exactly 1.0; how do I know that?) Similarly, a one unit increase in Z, holding fixed X, would reduce Y by about 0.62 units.

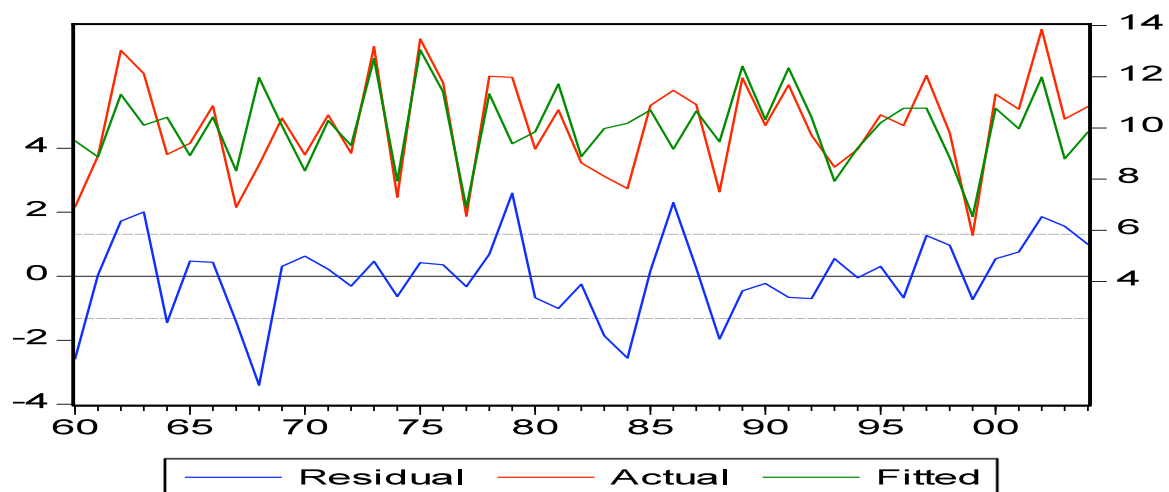
The Standard deviation (S.D.) of the dependent variable tells us about its variability around its mean value of 10.05. If the variable is normally distributed about 95% of the observations will fall within 2*S.D. of the mean, in this case between [6.17, 13.93]

R-square and adjusted R-squared are two measures of the in-sample “fit” of the regression equation. R-squared measures the proportion of the total variance of y that is “explained” by the

regression relationship (technically, it is 1 minus the ratio of the sum of squared errors to the variance of y). The adjusted R-squared takes the same calculation but scales up the SSR by $1/(T-k)$, where k is the number of parameters being estimated. This “penalizes” models with more parameters by bringing down their adjusted R-squared figures compared with more parsimonious models. In this case, both R-squared and adjusted R-squared indicate a moderate fit for this regression; an adjusted R-squared of .54 would be considered pretty good for a cross-sectional regression or for a time series regression of variables in “rate of change” form, but would be low for a time series model in levels.

The F-statistic is for a test that all parameters are jointly zero against the alternative that they are non-zero. The probability of the F-statistic clearly rejects that all params are zero. This is in some sense a summary test of the goodness of fit of the model.

The Durbin-Watson statistic is a test of the null hypothesis that the residuals from the regression are first order serially correlated. One needs to consult a table to be find out whether the statistic rejects the null. A value of 1.47 is far enough below 2 that we would want to check a table to see if we would conclude there is evidence of positive serial correlation. (I found a table online at <http://hadm.sph.sc.edu/courses/J716/Dw.html>) Turns out that this statistic is in the “indeterminate range” where we may or may not have a serial correlation problem. Looking at the plot of actual, predicted and residuals, below, one can see there does appear to be some modest serial correlation in the sense that positive values tend to be followed by positive values, and negative values by negative values, but it does not look very substantial.



9. (p. 31)

a. y_t , x_t , and z_t are variables; β_0 , β_1 , and β_2 are coefficients.

b. Units for β_0 are numbers of hot dogs. Units for β_1 are hot dogs per admission ticket; for β_2 , units are hot dogs per degree. The second two coefficients measure the responsiveness (formally the partial derivative) of hot dog sales to the particular variables.

c. Sign tells whether the relationship is positive or inverse. Sign on admissions is surely expected to be positive. I don't have strong feelings about the sign of the temperature coefficient; that is, I'm not sure whether people eat more or fewer hot dogs when it's hot. Maybe the coefficient is zero.

d. Taken rigidly, it's probably not sensible to allow a nonzero intercept. (Presumably hot dog sales must be zero if admissions are zero.) But more generally, if we view this linear model a merely a linear approximation to a potentially non-linear relationship, the intercept may well be non-zero (of either sign).

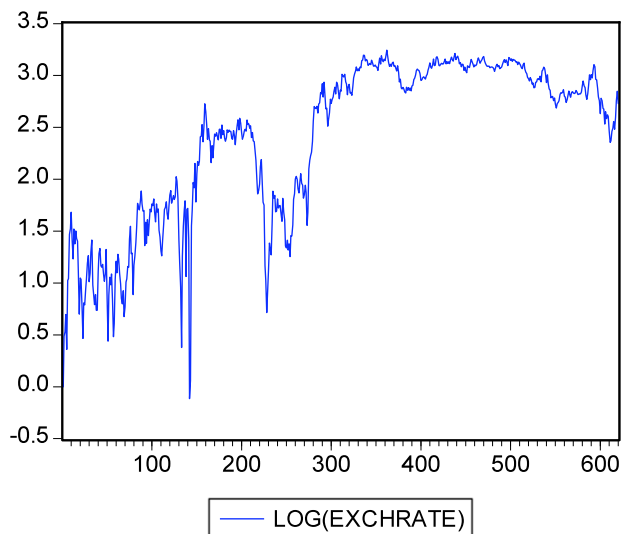
Ch 4. Problems and Complements:

6. (pp. 68-69) Graphical analysis of foreign exchange rate data using the "...\\Data Sets\\fcst4-02\\PC2_7_xyz.dat." data set.

- (a) Why might we be interested in examining data on the log rather than the level of the \$/Ft exchange rate?

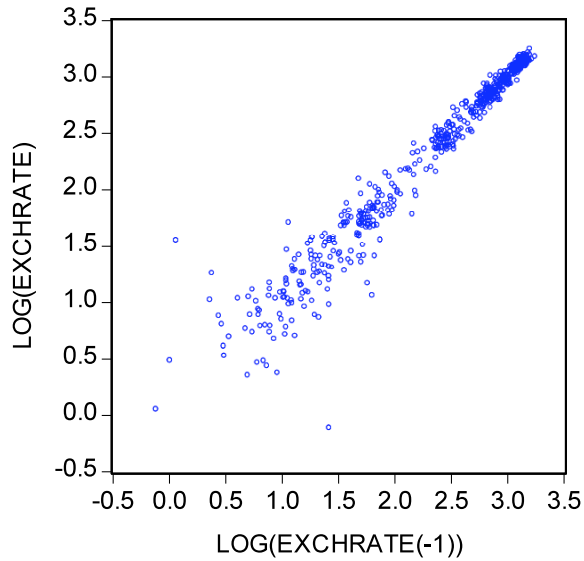
The log transformation tends to compress the extremes of a time series, which can produce a smoother graph. Also, the change in a log series over time is approximately the growth rate of a series, so one can sometimes see visually whether a series has a roughly constant growth rate or not.

- (b) Take logs and produce a time series plot of the log of the \$/Ft exchange rate. Discuss.



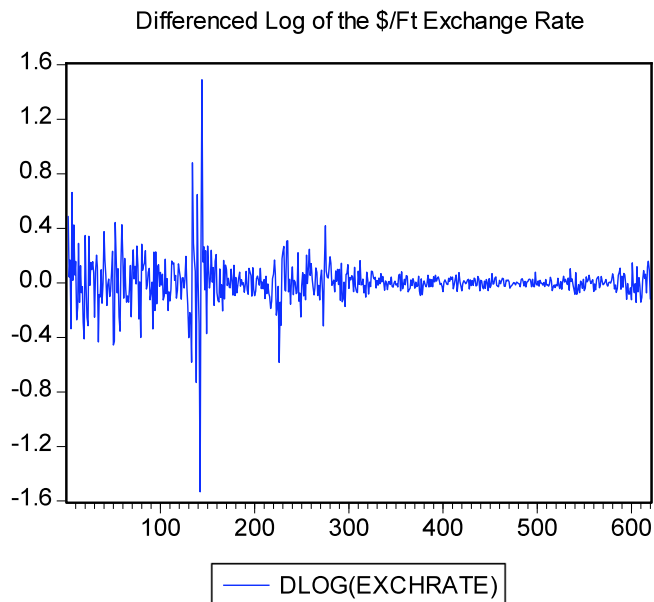
The Hungarian currency appreciated substantially during the first half of the 600 day period (but with substantial short-term swings along the way), and then had a fairly stable value against the dollar for the second half of the time period. There appears to have been much more volatility in the first half.

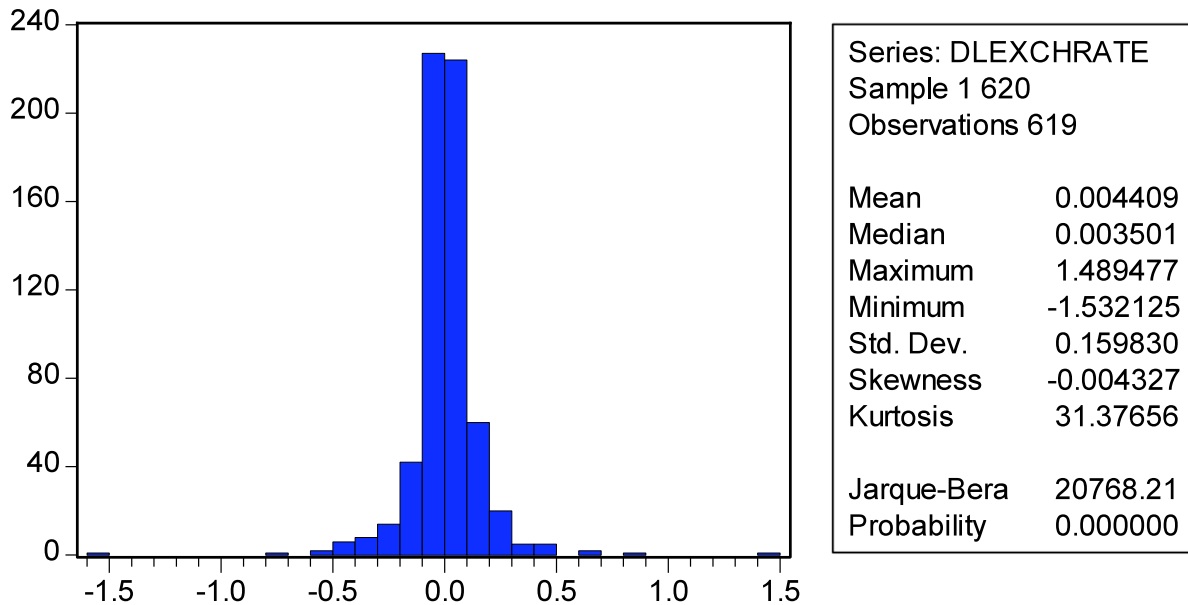
- (c) Produce a scatterplot of the log of the \$/Ft exchange rate against the lagged log of the \$/Ft exchange rate. Discuss.



You do this by graphing $\log(\text{exchrates})$ against $\log(\text{exchrates}(-1))$. What the figure tells us is that the log exchrates has a great deal of *persistence*, since there is a high correlation (tight scatter) between the variable and its value in the previous period. In fact, exchange rates are usually found to be very nearly a “random walk,” a term we will come back to later in the course.

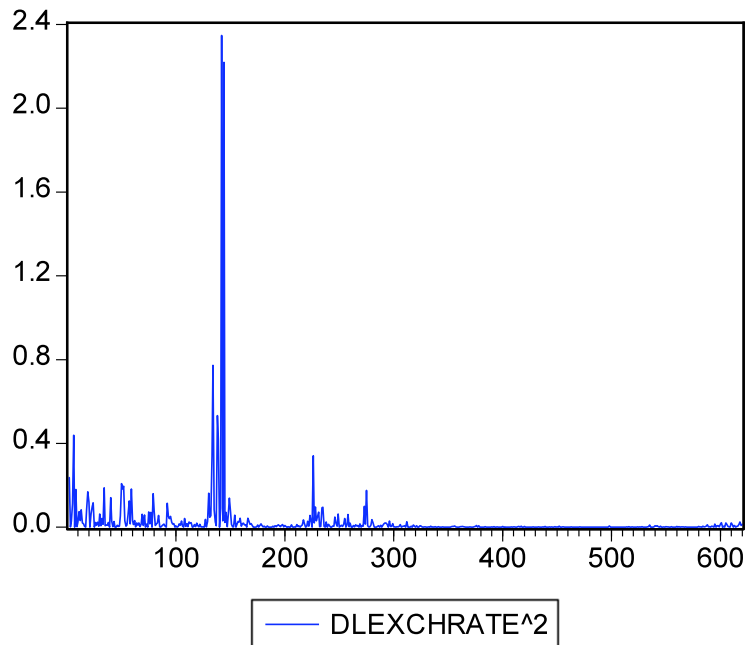
- (d) Produce a time series plot of the change in the log of the \$/Ft exchange rate, and also produce a histogram, normality test, and other descriptive statistics. Discuss. Do the log exchange rate changes appear normally distributed? If not, what is the nature of the deviation from normality? Why did we work with the differenced log rather than the original series?





From the time series plot of changes in the log exchange rate (the “differenced logs” of the series) it is clear that the transformed series is much closer to white noise than the original series: the dlogs clearly have a zero mean with no obvious autocorrelation. But it is also clear that the variability of the series is not constant (we’ll say more about this in part (e), below). From the histogram, we note that there are a lot of observations that are very small, but that there are also a significant number of positive and negative observations that are quite far from the mean—periods where the exchange rate moves by 1.5% or more in a single day! These are the periods in the time series graph where there are very large abrupt changes in the exchange rate. The Jarque-Bera statistic clearly indicates that the observations are not normally distributed, and that the source of the non-normality is high Kurtosis, that is “fat tails”—an unusually large number of observations at positive and negative extremes. (Normal distributions have a kurtosis about 3.)

- (e) Produce a time series plot of the square of the change in the log \$/Ft exchange rate. Discuss and compare to the earlier series of log changes. What do you conclude about the volatility of the exchange rate, as proxied by the squared log changes?



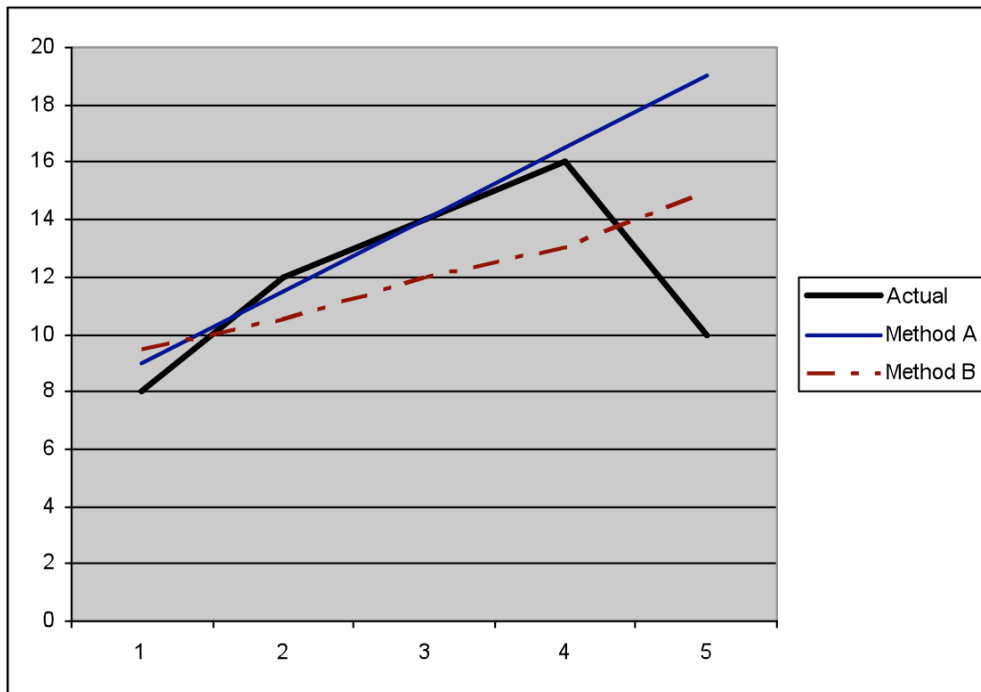
This graph of squared changes in the log exchange rate confirms that the variability of the series is not constant. Rather, there are periods of high volatility (especially in the first half of the sample) and others of very low volatility. Note that these squared changes are a rough measure of the variance of the log series.

Additional Problem:

In the first week of class, we looked at several measures of the goodness of forecast fit.

(a) Using the data below, calculate "mean absolute percent error" (MAPE) and "root mean squared percent error" (RMSPE) for both forecast series. (See textbook, page. 261-262.) Which model is better according to each measure? Why?

Period	Actual	Method A				Method B					
		Method A	Error	Abs Percent Error	Squared Perc Error	Method B	Error	Abs percent Error	Squared Perc Error		
1998	8	9	-1	0.13	0.02	8	9.5	-1.5	0.19	0.04	
1999	12	11.5	0.5	0.04	0.00	12	10.5	1.5	0.13	0.02	
2000	14	14	0	0.00	0.00	14	12	2	0.14	0.02	
2001	16	16.5	-0.5	0.03	0.00	16	13	3	0.19	0.04	
2002	10	19	-9	0.90	0.81	10	15	-5	0.50	0.25	
		MAPE=		0.22			MAPE=		0.23		
		RMSPE=			0.41			RMSPE=			0.27



Forecast A is preferred (barely) to forecast B according to the mean absolute percent error criteria, but forecast B is preferred to forecast A according to the root mean squared percent error criteria. Forecast A has the smallest average error, but makes one very large error in period 5. This one large error is severely penalized by the root mean squared percent error criteria, and so forecast B is preferred by the RMSPE criteria.

(b) Why might these percent error measures be more useful than mean absolute error or mean squared error?

The scale of mean absolute and mean squared error statistics varies with the scale of the data in question. By expressing the measures as percentages, the size of the errors is easier to gauge and to compare with forecasting errors typical in similar forecasting exercises. Note that both forecasts have fairly large average errors, in the 22-23% range.