

Lecture 20 **Time Series (Case Studies in Finance)**

STAT 102 Spring 2008

Part I

- We'll eventually look at two case studies
- Each case study involves the analysis of time series.
- Each involves a record of stock trading on NYSE
 - I Trading volume** (30 minute intervals)
 - II Daily returns**

Case I: Trading volume (30 minute intervals)

- **Trading volume** is the number of shares of stock traded in a given interval.
- Trading volume is known to be associated with volatility of a stock's share price.
 - The connection is (apparently) subtle and rather technical. SO
 - We will study data only about trading volume, and
 - We will *not* study data about the relation between volume and price.
- It's also thought there are technical connections between volume and price that can perhaps be profitably exploited.
- Volume is also of interest by itself in terms of understanding trading processes.
- We'll proceed in the spirit of understanding the basic structure of the volume process.

DATA:

- For today we will look at **one stock**, traded on NYSE
- This stock has the symbol **AIG** (= AIG Insurance Co)
- The data gives the number of shares traded during each 30 minute regular trading interval (generally 13 half-hour intervals each weekday, from 9:30am – 4pm)
- The period covered is 11/1/2005 through 2/28/2006

- Data on trading volumes is commercially available from several sources.
- However, some effort is required to process it in order to get a “clean” data-set.
- Our data was compiled by for us by a colleague from such a commercial source

Here are the first 16 lines of data:

Symbol	Date	Time (start)	T = interval #	Day of Week	# Trades	Vol/100
AIG	11/01/2005	:0:09:30:00	1	3	1065	13076
AIG	11/01/2005	:0:10:00:00	2	3	1111	11695
AIG	11/01/2005	:0:10:30:00	3	3	757	9186
AIG	11/01/2005	:0:11:00:00	4	3	571	5953
AIG	11/01/2005	:0:11:30:00	5	3	507	5287
AIG	11/01/2005	:0:12:00:00	6	3	460	4857
AIG	11/01/2005	:0:12:30:00	7	3	393	3482
AIG	11/01/2005	:0:13:00:00	8	3	346	4072
AIG	11/01/2005	:0:13:30:00	9	3	439	4599
AIG	11/01/2005	:0:14:00:00	10	3	886	10035
AIG	11/01/2005	:0:14:30:00	11	3	849	7686
AIG	11/01/2005	:0:15:00:00	12	3	713	6221
AIG	11/01/2005	:0:15:30:00	13	3	902	7050
AIG	11/02/2005	:0:09:30:00	1	4	700	4807
AIG	11/02/2005	:0:10:00:00	2	4	771	9409
AIG	11/02/2005	:0:10:30:00	3	4	746	10291

Variables in the Data

- The dependent variable is
 - Vol/100**: The volume of shares traded (in 100s) during the given half hour time interval.
- The other major variables recorded in the data are
 - T** = The time interval during the day: Coded as 1,..., 13.
 - Day of the Week**: Coded 2,..., 6.
 - Date**: 11/1/2005 – 2/28/2006;
but only days on which the Market is Open.
For convenience these are labeled in another column as
 - Date #**: Coded as 1, 2,..., 81.
- **# Trades** is another possible dependent variable, related to volume. We will ignore this variable (for now).

Goal of the Analysis

- “Understand” the process of Stock Volume
[*Subject to the available data*]
- Which covariates matter, and how do they affect volume?
 - How does the volume vary through the day?
 - How does the volume vary from day to day?
- How does the process behave in successive time periods
 - Are there any useful time-interdependencies (= auto-correlations) in the volumes?
- “Understanding” can also be measured via the ability to predict future volumes
 - Good predictions from one half hour to the next are potentially useful.
 - Predictions a day ahead may also be useful.

Model Creation

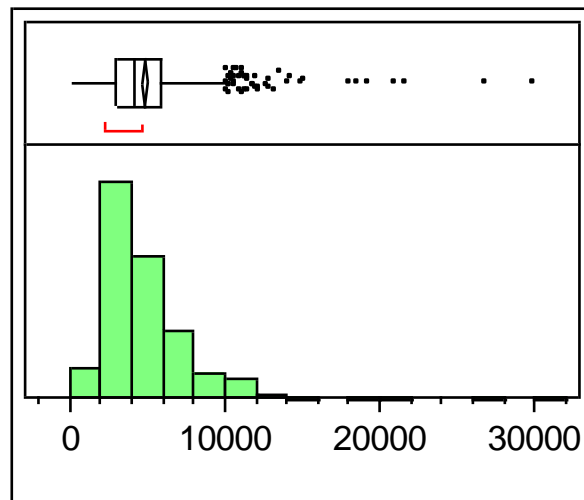
- What form of the dependent variable to use?
 - The variable itself or some transformation?
 - Can this be tentatively decided now? – If so, how to proceed?
 - Or should the decision be put off until later?
- Which covariates?
 - What model(s) can accommodate them?
 - Are any transformations needed?
- Time series structure:
 - What structure might exist?
 - How to build it?
 - How will it accommodate the covariates?
- **Build the Model** – requires many steps

Investigate the Proposed Model and Analysis

- Are there any outliers or other unusual data?
 - Do they need to be excluded?
 - Or, should they be included, but with special commentary?
 - Or, do they perhaps suggest changing the model?
- Do the remaining assumptions of the model seem satisfactory?
 - If not, do some modifications seem needed in the model?
 - Or, maybe this is the best possible analysis?
- Complete the analysis.
 - Report the results
 - Include some sample predictions
 - [— *Perform an out-of-sample validation if feasible*]

Interesting Plots and Tabular Output

- We will cooperatively model-build and analyze the data
- Here is some of the useful output that we will see in the process:



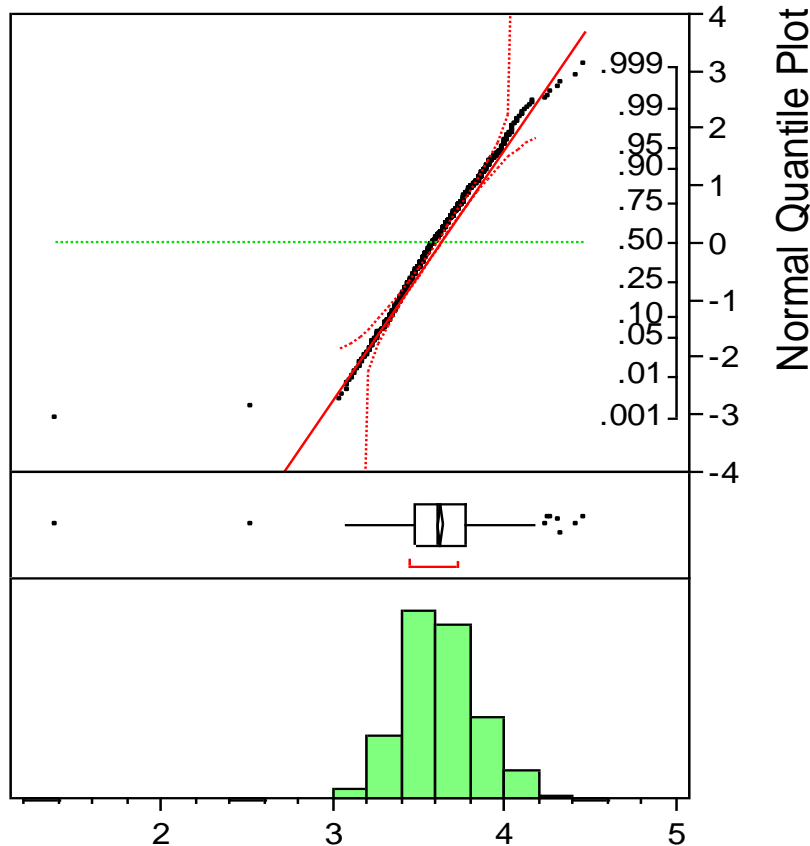
Histogram of Vol/100

Does this *suggest* that any transformation is needed?

Should we look at any other plots/tables related to this histogram?

Are there any individual data points to be concerned about?

For this plot, $N=1050$. Does that seem OK?



Does this histogram look “better” than the previous one?

Does that guarantee anything about future analyses? – Suggest anything?

Are there any individual data points to be concerned about here? Are they related to points we saw before?

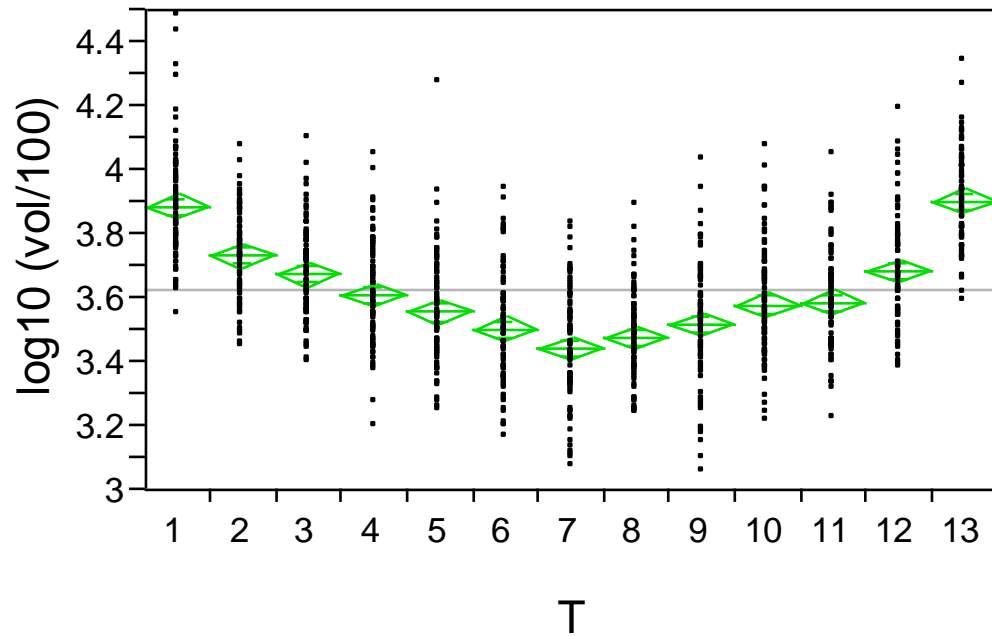
Should we do anything about them now?

Histogram of $\text{Log}(\text{Vol}/100)$

Answer to last question \Rightarrow go to table without 11/25/05. Use that for the remainder of the study.

“Seasonal” Variables

Are there any?



One-way ANOVA of $\text{Log}_{10}(\text{Vol}/100)$ vs $T = \text{time of day (in half hour intervals)}$

One-way ANOVA of $\text{Log}_{10}(\text{Vol}/100)$ vs day-of-week

Oneway Anova: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
day of week	4	0.569	0.142	3.081	0.0155
Error	1035	47.818	0.0462		
C. Total	1039	48.387			

Means for Oneway Anova

Level	Number	Mean	Std Error
2	169	3.616	0.0165
3	234	3.646	0.0141
4	221	3.657	0.0145
5	208	3.637	0.0149
6	208	3.591	0.0149

Comparisons for all pairs using Tukey-Kramer HSD

Level	Mean
4 A	3.657
3 A B	3.646
5 A B	3.637
2 A B	3.616
6 B	3.591

Lessons from the Previous One-way Analyses?

- Time of day matters.
- Day of the week seems to matter a little.
- Can we use BOTH of these factors? If so, HOW? What happens if we try to do so?
- Here are some relevant tables and plots:

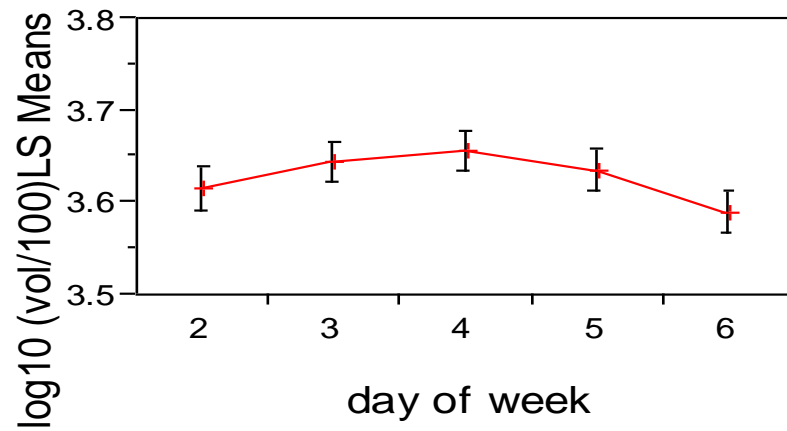
What are they and what do they show?

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	16	20.911	1.307	48.66
Error	1023	27.476	0.027	Prob > F
C. Total	1039	48.387		<.0001

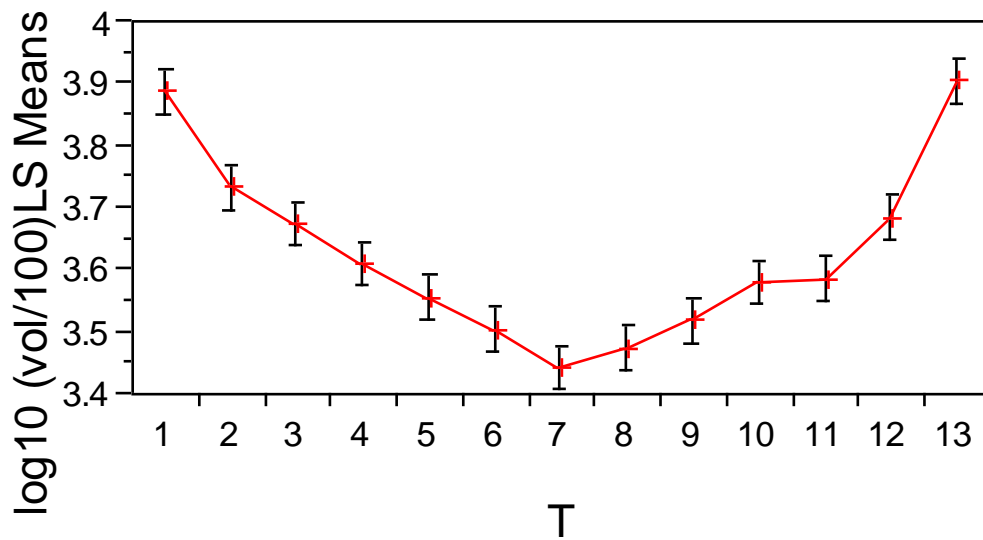
Effect Tests				
Source	DF	Sum of Squares	F Ratio	Prob > F
day of week	4	0.569	5.300	0.0003
T	12	20.342	63.115	<.0001

Summary of Fit	
RSquare	0.4322
Observations	1040

LS Means Plot



LS Means Plot



Is there an overall time trend?

- Since both T and Day-of-week seem to matter (“a lot” and “some”, resp.) should we look for a trend after controlling for them? How?
- Here is the key table from an analysis of $\text{Log}_{10}(\text{Vol}/100)$
- **What analysis is it from and what does it show?**

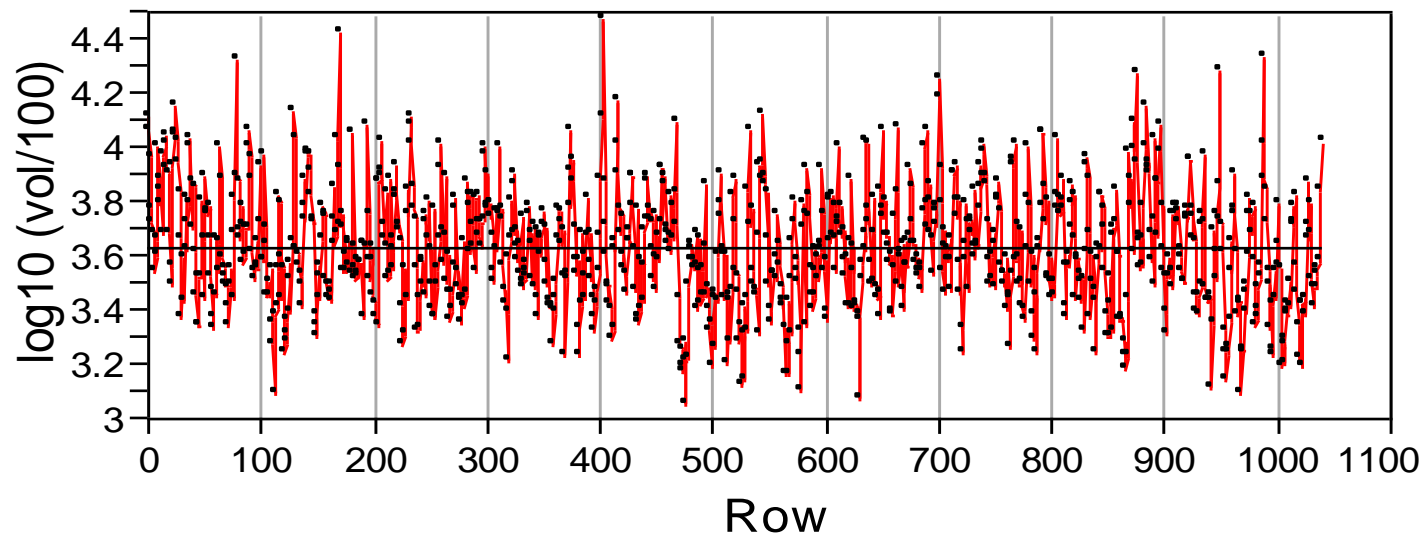
Effect Tests				
Source	DF	Sum of Squares	F Ratio	Prob > F
day of week	4	0.534	5.05	0.0005
T	12	20.342	64.10	<.0001
Date #	1	0.447	16.91	<.0001

Exercise: What is the value of R^2 for this analysis?

- Approximately what?
- Exactly what? [The JMP table shows $R^2=44.1\%$.]
- Is the coefficient for Date# positive or negative? (Can't tell without using JMP.)
- What would have happened if we hadn't controlled for T and day-of-week? (Subtle, but can correctly guess w/out JMP.)

Time Series Effects

- Are there time-series effects?
- Shouldn't we have taken a first look before now? — Probably!



Time-series plot w/out controlling on seasonal effects and trend.

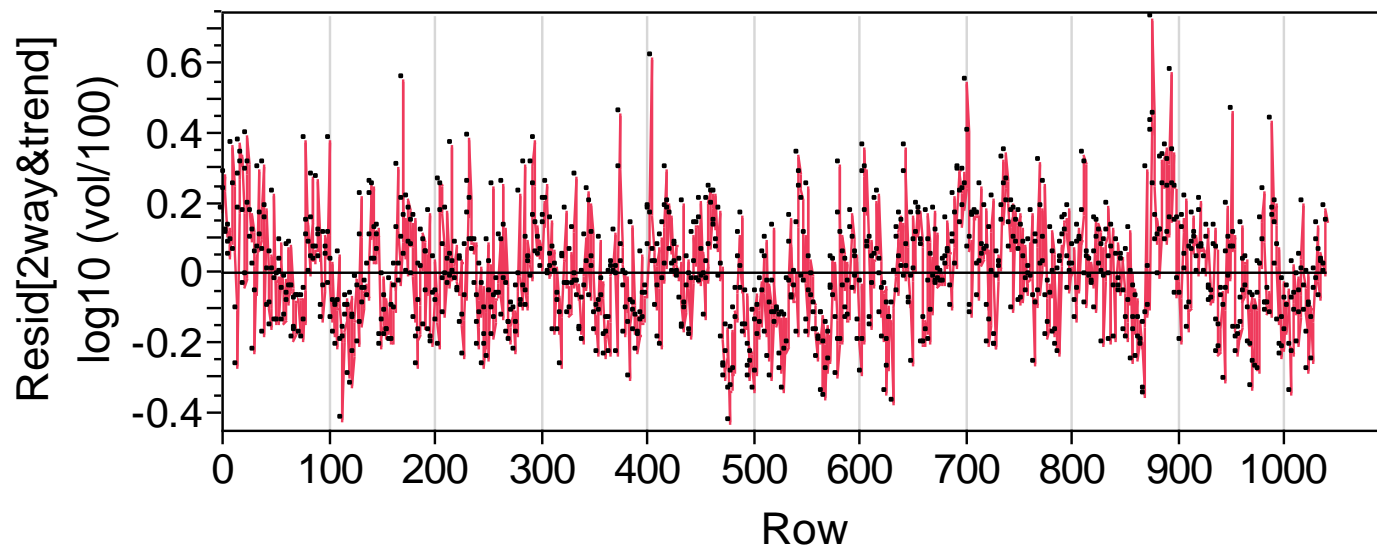
Does this look like a pure time series?

[It's hard to see why not! - Even though we now know it's not]

- The plot is so full of points that
- It's also hard to see that this time series has much of an AR(1) structure, (either with or without controlling for trend and seasonality)
- But it does!
- To check for the existence of this AR effect it's easiest to examine the residuals from the previous analysis.
- IF there is an AR effect, the best analysis is to put the appropriate lagged variable into the model, and re-analyze it.

Checking the Residuals for Presence of an AR Effect

- Save the residuals from the previous analysis involving T, day-of-week, and date#.
- Apply the JMP “time series” analysis. Here is the plot



- This plot has stretches of + values followed by stretches of – values. This suggests presence of a time series correlation.

- To use the time series platform to test for autocorrelation of the residuals Use the “ARIMA” button with “autoregressive order” = 1. Output is –

Parameter Estimates					
Term	Lag	Estimate	Std Error	t Ratio	Prob> t
AR1	1	0.5338	0.0262	20.36	<.0001
Intercept	0	0.000348	0.00198	0.18	0.8603

- Note that the AR1 coefficient is **very significant**.
- Because of this we should conduct a complete analysis that incorporates an AR term.

Construct a Complete Analysis

- Create a column with Lag1 values of $\log_{10}(\text{vol}/100)$
- Run a FitModel with ModelEffects: T, day-of-week, Date# and Lag($\log_{10}(\text{vol}/100)$). Here is output:

Summary of Fit

RSquare	0.5998
Root Mean Square Error	0.1374
Observations	1039

Effect Tests

Source	DF	Sum of Squares	F Ratio	Prob > F
T	12	8.327	36.73	<.0001
day of week	4	0.142	1.87	0.1126
Date #	1	0.085	4.51	0.0339
Lag(Log10(vol/100))	1	7.729	409.15	<.0001

- The **AR(1)** effect is highly significant; with **P<.0001**.
- R^2 has increased from 44.1% without the AR(1) term

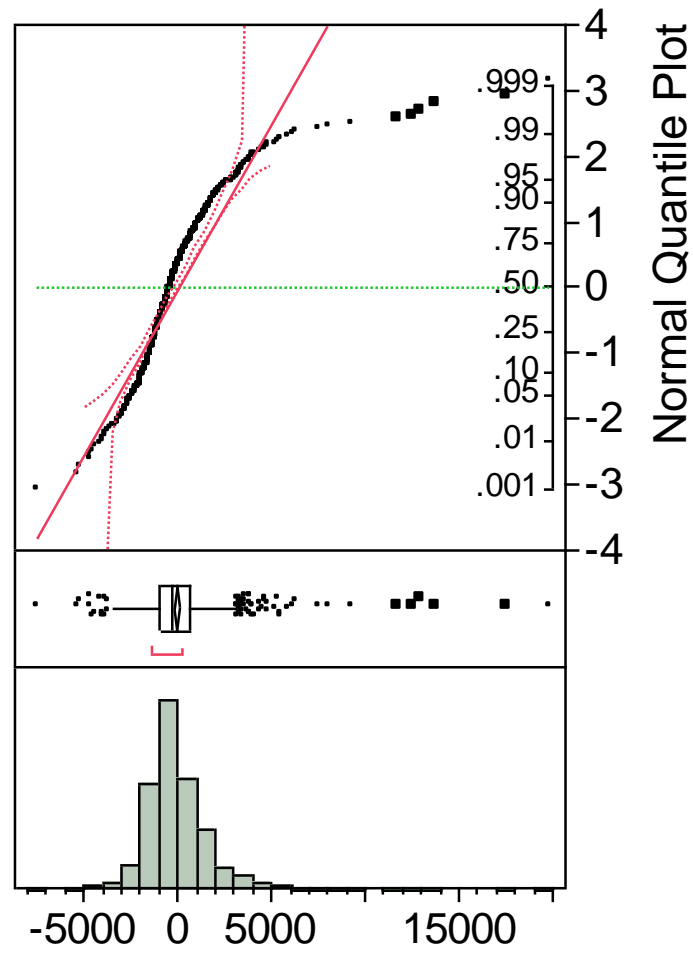
- The AR(1) analysis is clearly preferable to the one without an AR(1) effect.
- Why is the AR(1) effect so big?
There's a very good stock-market explanation for this to be a big effect. Does anyone know what the explanation is?
- Note (as a curiosity) that the day-of-week is no longer statistically significant.
- Forecasts can be obtained by the method described in the previous lecture.
- For forecasting $\log_{10}(\text{vol}/100)$ for T[1] , on date# =82, which is a day-of-week=4, we have

$$\text{Forecast} = 1.707 + .110 + .016 + (-.000392) \times 82 + .534 \times 3.849 = 3.920$$

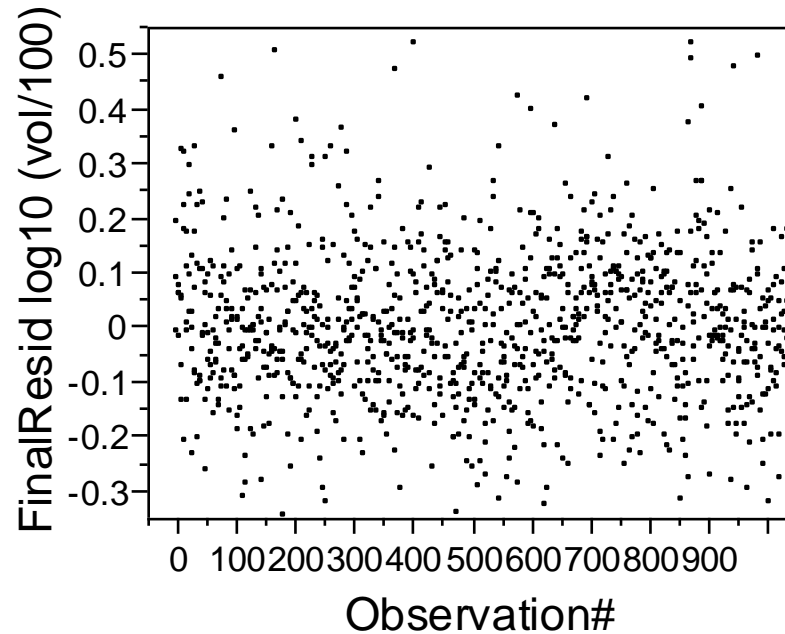
(See JMP data table – not given - for coefficients)

Residuals and other Model Diagnostics

- We need to look for:
 1. Unwanted outliers – that have not already been removed
 2. Heteroscedasticity – variance of residuals non-constant, and depends on some covariate
 3. Non-normality of residuals
- The easiest plot to get and diagnose is the histogram and normal residual plot of residuals (see next page):
 - That plot looks not terrible even though there are about 15 overly large residuals (out of over 1000).
 - Only **5** are terribly large, but they don't seem to be “outliers” that should be excluded.
 - They involve isolated large volumes. This is unfortunate, since those unusually large volumes may be the most important to predict. But it seems they can't be well predicted using the type of analyses here.



Scatterplot of residuals vs time order



This seems satisfactory; it shows no fan-shaped pattern or other sign of irregularity. (You can see the same slight right skewness shown in the histogram; but that's not heteroscedasticity.)

Is There Heteroscedasticity by Time of Day?

Such heteroscedasticity would show up in dramatically varying standard deviations for the residuals at different times of day. But the following table of standard deviations (taken from within the “Fit Y by X” platform) shows these SDs to be similar in size; even though the difference may be statistically significant.

Means and Std Deviations		
Level	Number	Std Dev
1	79	0.1679
2	80	0.1163
3	80	0.1234
4	80	0.1260
5	80	0.1391
6	80	0.1471
7	80	0.1328
8	80	0.1213
9	80	0.1670
10	80	0.1451
11	80	0.1241
12	80	0.1413
13	80	0.1176

Part II Daily returns

Volatility in Financial Time Series

Data: Daily log returns for IBM stock for a 35+ year period.
(1970-mid 2005)

[The “return” here is based on the daily opening per-share PRICE of the stock (adjusted for splits, and excluding dividends).

The formula for Return is

$$\text{Return}_t = \text{PRICE}_t / \text{PRICE}_{t-1}.$$

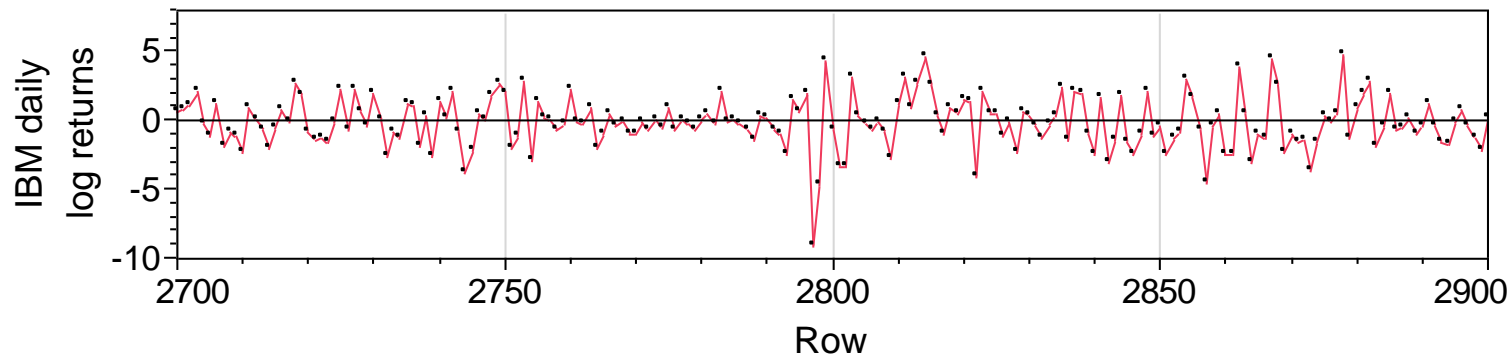
Log Return here means

$$\text{LogReturn}_t = 100 \times \text{Log}_e(\text{Return}_t).$$

[The multiplicative factor “100” in this expression is intended as a convenience in order to avoid small decimals.]

Time Series Plot and Analysis

- The entire time series plot has 8938 data points. That's too many to see the time series pattern clearly. So here is a typical segment of the plot – from day 2700 to day 2900.



- This looks **almost** like a pure time series. In particular, there is
 - no apparent trend
 - no apparent seasonal effect
 - no apparent autocorrelation

- To check, lets do an analysis with possible trend and AR(1) effects.
- *(We don't have day-of-week labeled, so we can't really condition on day-of-week, but there doesn't appear to be any 5 day"seasonal" effect in the time series plot.)*

RSquare	.000007252
Root Mean Square Error	1.481035
Observations	8937

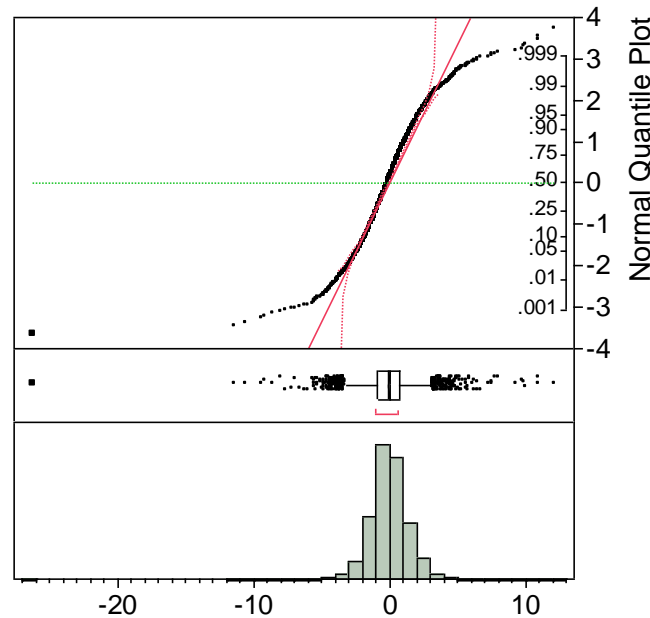
Term	Parameter Estimates			
	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.0450	0.0313	1.44	0.1509
Day	-1.303e-6	6.073e-6	-0.21	0.8301
lag1 LogReturns	0.00144	0.0106	0.14	0.8915

- There's **No Evidence** here of either a linear trend or of an AR(1) effect.

Nevertheless

- There's something "not right" about the time series
- It's visible in the segment of the time series plot we've shown above. (It's even more visible in some other segments of the plot.)
- Can you see it???
- It's also visible in the histogram of residuals from the preceding analysis:

Histogram of Residuals:



- There's one very noticeable (negative) extreme value (*Outlier?*)
- More important, the distribution is not really normal –
- It has heavy tails in both directions!
- There is a better explanation than just, “**OOPS!**”

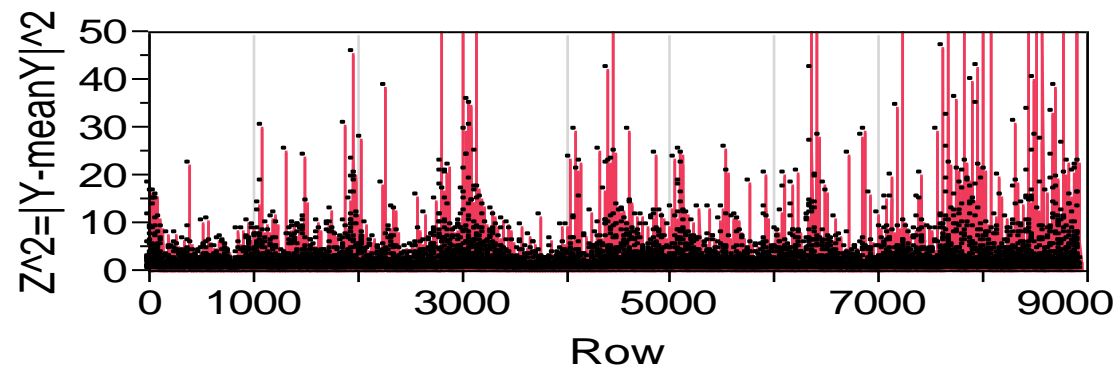
Persistent Volatility

- Create a column for the squared residuals from their mean of DailyLogReturns.
- [So, if Y_i denotes the DailyLogReturns then the new column is

$$Z_i^2 = (Y_i - \bar{Y})^2 .]$$

- Now run an ordinary AR(1) analysis of the Z_i^2 variables.
 - In order to do so you need to create a column of lag1 values of Z_i^2 .
 - We should not expect the Z_i^2 variables to be normally distributed. So, none of the usual hypothesis tests in this new analysis need to be valid. But the analysis is informative anyway –
- Here is a plot of the Z_i^2 :

Time Series Plot of Z_i^2



- This doesn't look like a typical AR(1) pattern.
- There are many spikes
 - Some of them occur singly
 - **But many times they appear to be grouped together**
 - Their autocorrelation is 0.159
 - This is not large, but it's definitely not 0.

(A valid statistical test would justify that conclusion.)

Brief Explanation

Volatility is the variance of \log_e Return of an asset. (Our Z_i^2 is the daily volatility.) Volatility is of fundamental importance in empirical finance. Portfolio theory, asset pricing and hedging all involve volatility.

The above analysis indicates that volatility for IBM stock returns is not constant; **AND** the current volatility depends on the previous volatility, with periods of high volatility tending to follow each other and periods of low volatility tending to follow each other.

“Conditional heteroscedastic models” are probabilistic models for modeling the volatility of an asset’s returns. Robert Engle won the Nobel Prize in Economics (04) for his development of an important conditional heteroscedastic model for volatility called the autoregressive conditional heteroscedastic model (ARCH).

For much more information about time series, you should consider taking

Statistics 434

Financial and Economic Time Series, which will be offered next fall.