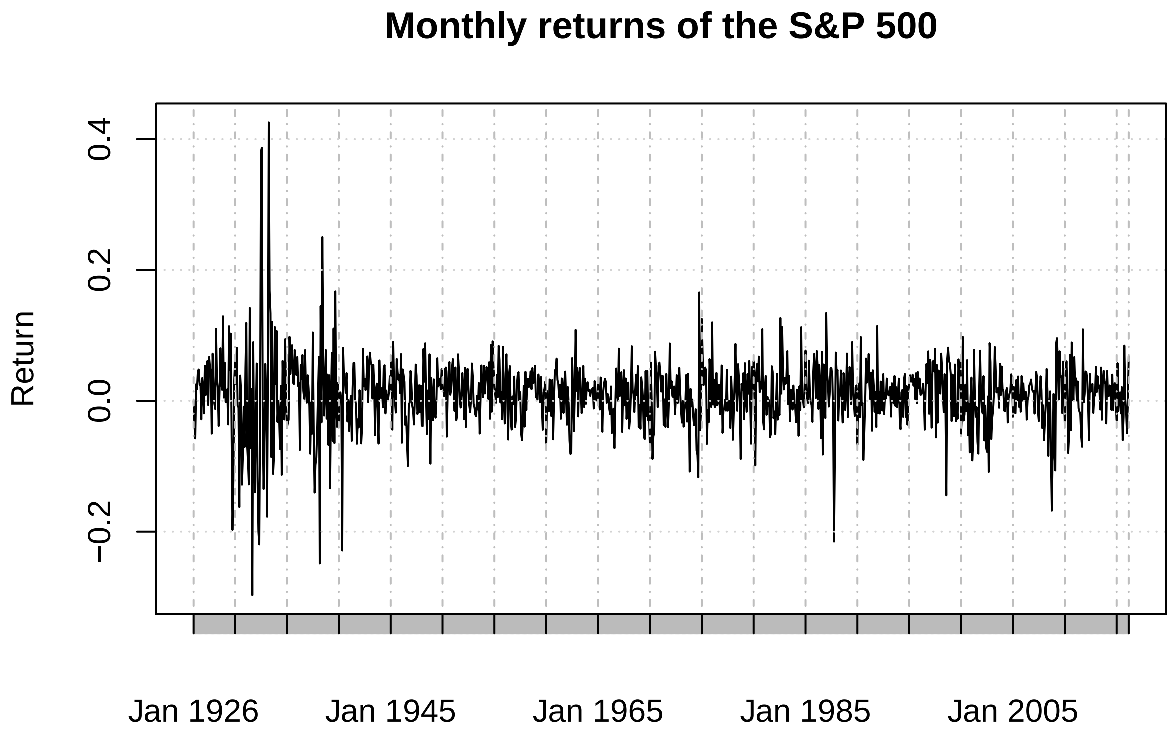
Why be normal?

Rex Macey

May 8, 2016

This document is a follow-up to the Hultstrom Hypothesis test. In this, I look for distributions that fit our monthly S&P 500 returns better than the normal distribution. Let's start with a summary of the returns.



```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.297300 -0.016720  0.012950  0.009447  0.039000  0.425600
```

The standard deviation is 0.054389.

The Distributions

We fit three distributions: normal, lognormal and Student t. The parameters for the normal distribution are: Mean: 0.009447

SD: 0.054364

A variable is said to be lognormally distributed if its log is normally distributed. In finance this is often used. It's the distribution underlying the Black-Scholes option pricing model. In this case, it is assumed that 1+r is lognormally distributed, so we add one to each of the returns. The parameters we find are: Meanlog: 0.007946

SDlog: 0.054162

Last but not least we use a Student t distribution. This is often used to test for significance when normality cannot be assumed. The distribution has fatter tails than the normal distribution which is why it was chosen. The Student t distribution approaches a normal distribution when the degrees of freedom (df) is high. Here we find the df parameter which is: df: 9.127143

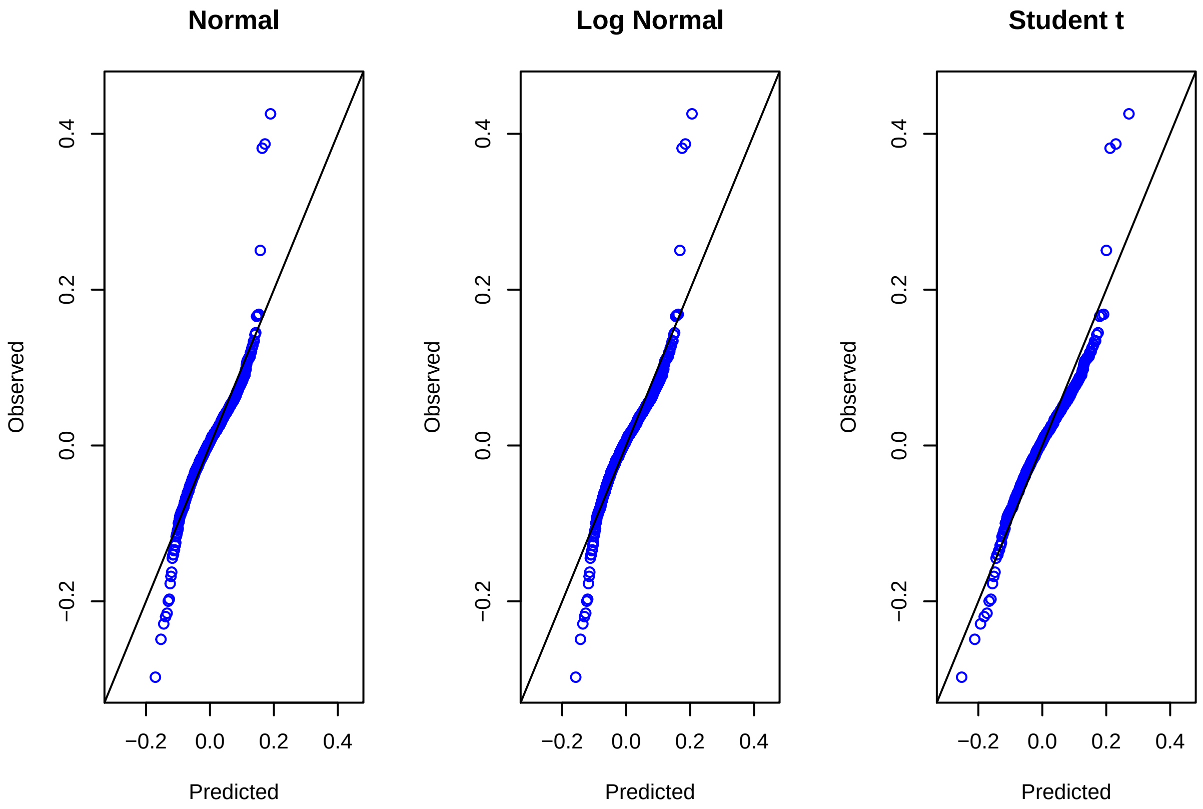
The Fits

For each of the three models we will look at the R-Square, the root mean square error, and the mean absolute error. In addition we compare a summary of the distribution with the observed returns.

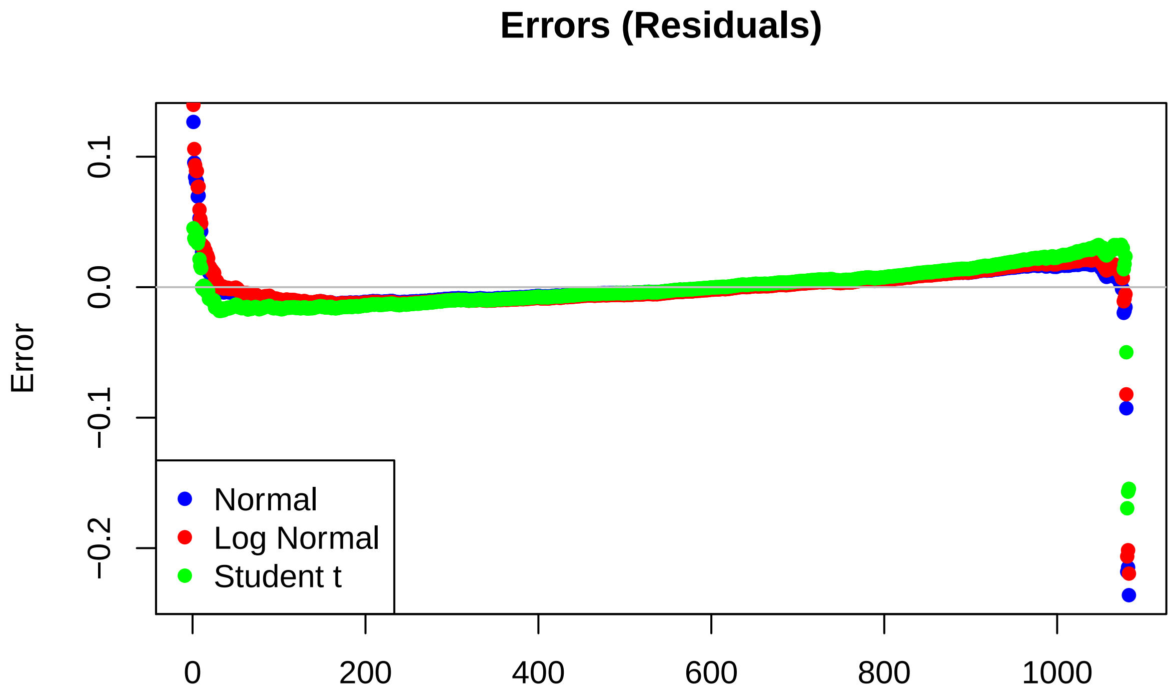
	R2	RMSE	MAE
Normal	0.901	0.562	0.010
Log Normal	0.898	0.573	0.010
Student t	0.913	0.528	0.012

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Observed	-0.2973	-0.01672	0.012950	0.009447	0.03900	0.4256
Normal	-0.1707	-0.02718	0.009447	0.009447	0.04608	0.1895
Log Normal	-0.1576	-0.02814	0.007977	0.009455	0.04544	0.2061
Student t	-0.2521	-0.02871	0.009447	0.009447	0.04760	0.2710

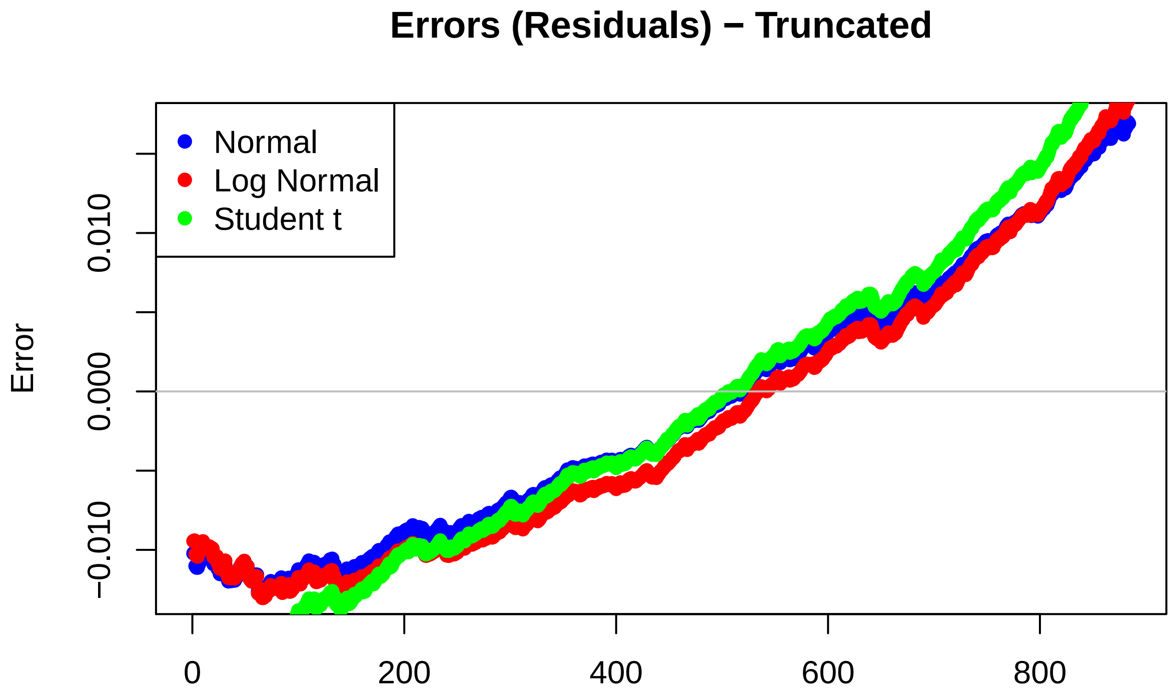
The observed (actual) returns have a wider range (Min to Max) than the predicted, but the Student t comes closer and has better R2 and RMSE values. However, the normal fit performs better on mean absolute error. Below are plots of the predicted v the observed returns. The Student t does come closer on the extremes, particularly the downside.



The picture isn't entirely clear as the following chart shows. While the Student t (green) has smaller errors at the extreme, it has larger errors in other regions. That is why the MAE score is higher.



To see the errors toward the center more clearly, the chart below truncates the 100 most extreme returns on each side.



Conclusion

I don't see an unequivocable best fit. If one is more concerned about capturing the extremes, one would use the Student t. If one cares equally about all predictions and if errors are penalized by a factor of 1 (meaning a 0.03 error is 3x worse than a 0.01 error), then the normal distribution is better.

There might be a Frankenstein way to combine these in a piece-wise fashion. We might use the Student t to predict extremes and the normal for non-extreme. This may not be crazy if one believes that returns in a crisis come from a different distribution than returns in a "normal" market. If one were to make such an assumption, then one would could improve the fit in a normal market where from the preceding chart we observe a pattern in the errors which could be corrected.