David Smith, Madison Melusen, Emily Lin

A Time Series Model for Gold Prices

Introduction

The purpose of this report is to predict gold prices in the year 2009 in US Dollars. The price of gold was collected monthly from January 2001 to December 2008 and was recorded. We made a time series model for monthly gold prices for this time period by analyzing the data to ultimately obtain the best model to predict future gold prices. In order to do this, we looked at the plot to determine what kind of trend we had, determined if we had to transform the data to account for anything, and determined a preliminary model. Then, we checked for assumptions and tested for the existence of autocorrelation, and reevaluted our model. We conclude with a discussion of the usefulness of our model for predicting future gold prices.

Data Summary

The explanatory variable in this project is the month and year (referred to as time), while the response variable is the price of the gold (measured in US Dollars, referred to as price). The trend is positive since the price of gold increases over time. The data appears to curve upwards, but it was unclear if it was a growth or quadratic trend. (see Original Plot, Appendix). To choose the best model we looked at the standard error and adjusted r squared of both models. Using the proc reg procedure in SAS, the growth curve model had a standard error of e^0.075 = 1.078 and an adjusted r squared value of 0.963, whereas the quadratic model had a standard error of 44.712 and an adjusted r squared value of 0.95. Since the growth curve model had a higher adjusted r

squared and a *much* lower standard error, we chose to model our data with a growth curve. Plotting the growth curve model using a natural logarithm transformation, we saw that it now appeared linear (See Growth Model Plot, Appendix).

We tested for possible seasonal variation with the dummy variable method, and the p-value was 1, an extremely high value, indicating that there was no seasonal variation. We also tried using trigonometric functions to account for seasonal variation, but again, the results were not significant. We transformed the data using the natural logarithm to handle the growth curve model, and ended up with a hypothesized model of $y^*(t) = \ln(y(t)) = \alpha_0 + \alpha_1 t + u_t$, where $y(t) = \beta_0\beta_1{}^t\varepsilon_t$.

Analysis

Our hypothesized model was $y^*(t) = \ln(y(t)) = \alpha_0 + \alpha_1 t + u_t$, where $y(t) = \beta_0\beta_1{}^t\varepsilon_t$. In order to decide whether or not to keep this model, we had to check some assumptions.

According to the Residual QQ Plot and the Residual Plot (see Residual Plots, Appendix), no regression assumptions were severely violated. The error terms had an approximately normal distribution, were centered around 0, and had approximately constant variance (although there was a relative fanning out pattern).

In order to check for autocorrelation, we used the Durbin-Watson test. The p-value for the Durbin-Watson statistic for positive autocorrelation was almost 0, suggesting the existence of positive first-order autocorrelation. To handle this, we rewrote our model as $y^*(t) = \ln(y(t)) = \alpha_0 + \alpha_1 t + (\Phi u_{t-1} + a_t)$, with $(\Phi u_{t-1} + a_t)$ having replaced u_t, and used the proc arima procedure in SAS to reevaluate our model. We found that all of the variables were

significant, with p-values close to 0. The standard error was now $e^{0.04} = 1.004$, which was even smaller than before.

The final model that is used is $y^*(t) = \ln(y(t)) = \alpha_0 + \alpha_1 t + (\Phi u_{t-1} + a_t)$, where $y(t) = \beta_0 \beta_1^t \varepsilon_t$. Since the variance of the error terms is not perfectly constant, we should be careful when using this model for prediction, however we can assume that this model will be generally accurate in prediction.

Results/Conclusions

Our prediction equation is $\hat{y}^*(t) = \ln(\hat{y}(t)) = 5.544 + 0.012t + 0.882\hat{u}_{t-1}$ (see Arima Output: Growth Curve Model). At time t = 0 (December 2007), the gold price is expected to be $e^{5.544} =$ $255.70. This is the interpretation of the intercept. The growth rate is $e^{0.012} - 1 = 1.21\%$. In general, the gold price is expected to increase exponentially with each passing month, given our model.

The predicted intervals for gold prices in 2009 are given in Table of 2009 Prices, Appendix. The width of the intervals increases as time increases, due to the exponential growth and increasing variation of the model, which is a problem for accurately predicting future gold prices. Comparing our model to actual gold prices in the last ten years, we find that gold prices dropped significantly from 2012 to 2015, and have remained much lower than their peak in 2011. Thus, our model turned out to be inaccurate. Intuitively, this can be explained by the fact that gold prices cannot increase exponentially indefinitely in the real world.

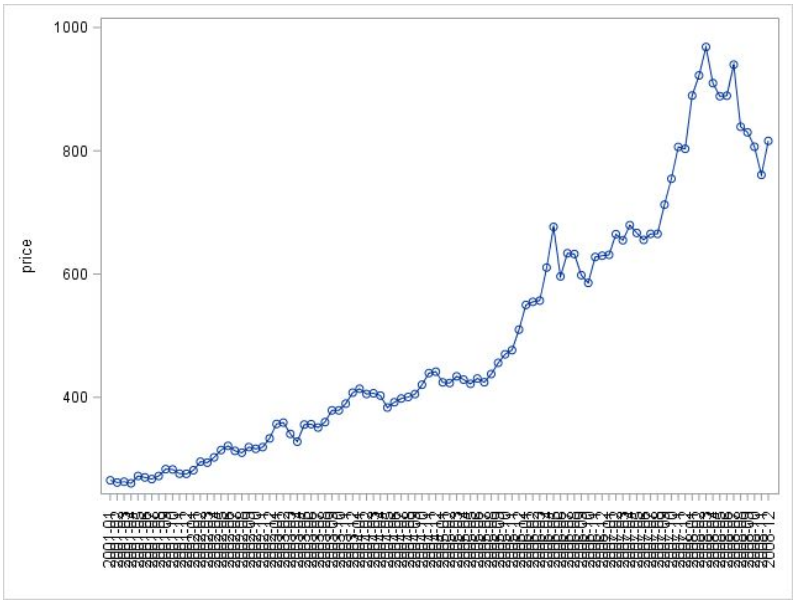
Areas of future research might be to rerun this analysis with more recent data (for instance, 2009-2017). The real data for the price of gold does not increase infinitely as our model

predicts, so using the recent data can help us create a more accurate model for predicting prices in the future (i.e. 2018 and beyond). The researcher should be aware that the regression assumptions (especially the assumption of constant variance) was not perfectly satisfied and, therefore, one should be careful when using this model to make predictions. To improve the research design, one can use more frequent data, if it can be reasonably collected. For example, instead of using the average gold price for each month, one can use the average gold price for each week. That way, we can build an even more accurate model for the prices.

Appendix

```
*Plot original price over time;
proc sgplot data=gold sganno=sganno pad=(bottom=15%)
series x=month y=price /markers;
xaxis display=(nolabel novalues);
Run;
```

Original Plot

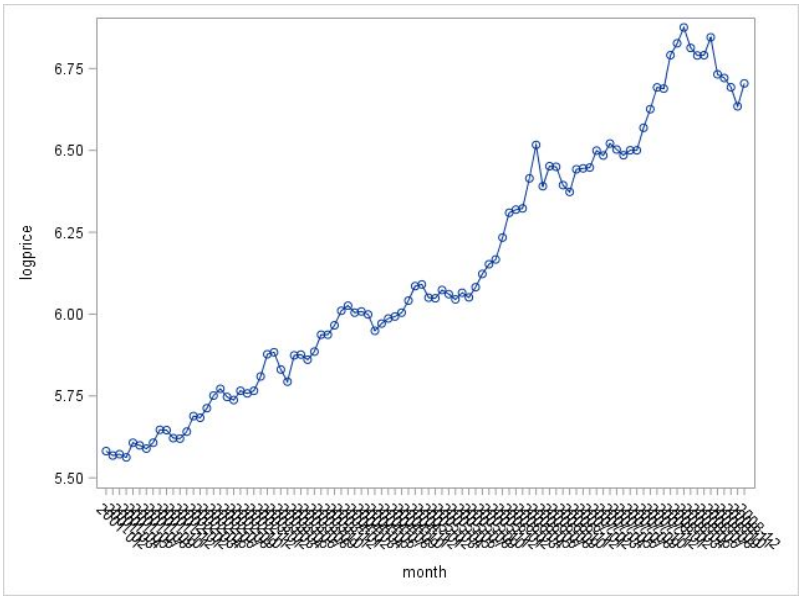



```
*Create new data set in preparation for transformations;
data gold2;
set gold;
```

```
time = _n_;
logprice = log(price);
time2 = time**2;
Run;

*plot growth curve model;
proc sgplot data=gold2;
series x=time y=logprice / markers;
run;
```

Growth Model Plot




```
*Evaluate the growth curve model;
proc reg data=gold2 plots=none;
model logprice = time;
run;

*Evaluate the quadratic model;
proc reg data=gold2 plots=none;
model price = time time2;
run;

*Test for autocorrelation;
proc reg data = gold2 plots=none;
model logprice = time /dwprob;
run;
proc reg data = gold2 plots=none;
model price = time time2 /dwprob;
Run;
```

DWPROB Output, Growth Curve Model

The REG Procedure
Model: MODEL1
Dependent Variable: logprice

Durbin-Watson D	0.294
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	96
1st Order Autocorrelation	0.839

DWPROB Output, Quadratic Model

The REG Procedure
Model: MODEL1
Dependent Variable: price

Durbin-Watson D	0.357
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	96
1st Order Autocorrelation	0.792

```
*Compare growth curve model and quadratic model;
proc arima data = gold2 plots = none;
identify var = logprice crosscor = time noprint;
estimate input = time p = (1);
run;
proc arima data = gold2 plots = none;
identify var = price crosscor = (time time2) noprint;
estimate input = (time time2) p = (1);
run;
```

Arima Output: Growth Curve Model

The ARIMA Procedure

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	5.54429	0.03669	151.12	<.0001	0	logprice	0
AR1,1	0.88153	0.05018	17.57	<.0001	1	logprice	0
NUM1	0.01240	0.0007683	16.14	<.0001	0	time	0

Constant Estimate	0.65682
Variance Estimate	0.00158
Std Error Estimate	0.039743
AIC	-343.873
SBC	-336.18
Number of Residuals	96

```
*Check for seasonal variation using dummy variables;
%macro dumvar;
data gold3;
set gold2;
month2 = SUBSTR(month,6,2)*1;
%do i=1 %to 11;
Mdum&i. = 0; if month2=&i. then Mdum&i.=1;
%end;
run;
%mend;
%dumvar;

*Model with dummy variable seasonal variation;
proc reg data=gold3 plots=none;
model logprice = time Mdum1-Mdum11;
Run;



*Check for seasonal variation using trigonometric equations;
data gold4;
set gold3;
time=_n_;
pi = arcos(-1);
two = (pi*time)/6;
sintwo = sin(two);
costwo = cos(two);
```

```
run;

*Model with trigonometric seasonal variation;
proc reg data=gold4 plots=none;
model price = time sintwo costwo;
Run;

*Predict 2009;
proc arima data = gold2 plots=none;
identify var = logprice crosscor = time noprint;
estimate input = time p=(1);
forecast printall out=goldpred;
run;
Quit;
```

Table of 2009 Prices

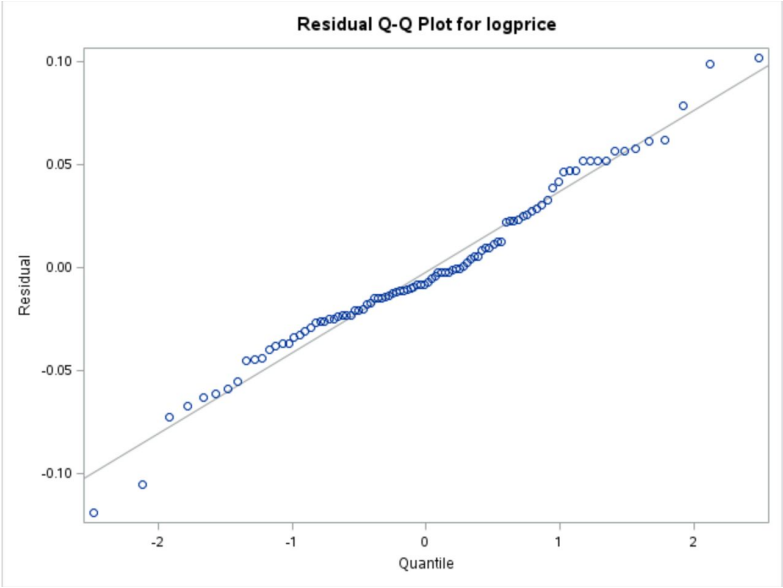
Month	95% confidence limits
January 2009	($767.16, $896.50)
February 2009	($759.22, $934.48)
March 2009	($758.39, 964.49)
April 2009	($761.05, 990.09)
May 2009	($766.01, 1012.82)
June 2009	($772.39, 1033.70)
July 2009	($779.84, 1053.10)
August 2009	($788.08, 1071.48)
September 2009	($796.87, 1089.09)
October 2009	($806.09, 1106.21)
November 2009	($815.74, 1122.93)
December 2009	($825.75, 1139.22)

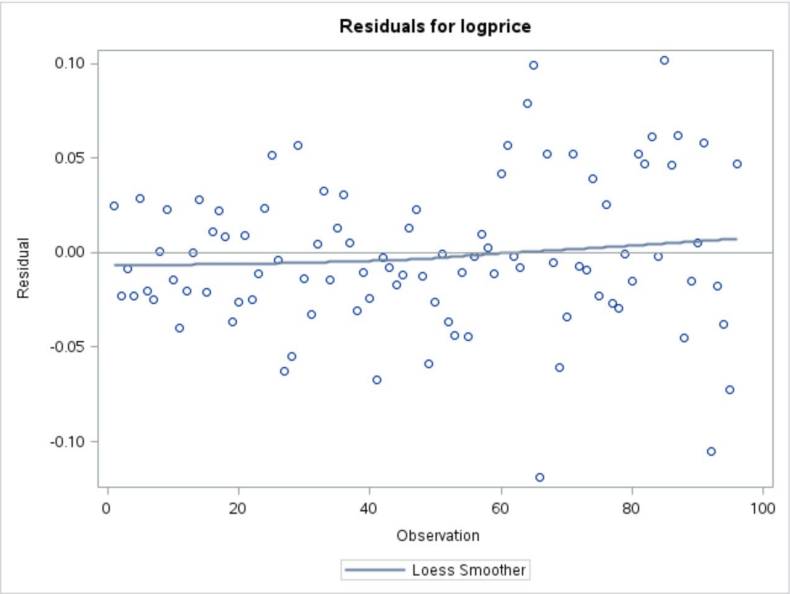
SAS Output: Predicted Values

97	6.7206	0.0397	6.6427	6.7985	.	.
98	6.7362	0.0530	6.6323	6.8400	.	.
99	6.7514	0.0613	6.6312	6.8716	.	.
100	6.7663	0.0671	6.6347	6.8978	.	.
101	6.7808	0.0713	6.6412	6.9205	.	.
102	6.7952	0.0743	6.6495	6.9409	.	.
103	6.8093	0.0766	6.6591	6.9595	.	.
104	6.8232	0.0784	6.6696	6.9768	.	.
105	6.8369	0.0797	6.6807	6.9931	.	.
106	6.8505	0.0807	6.6922	7.0087	.	.
107	6.8639	0.0815	6.7041	7.0237	.	.
108	6.8772	0.0821	6.7163	7.0381	.	.

*Check regression assumptions;
proc arima data = gold2 plots(only)=Residual(QQ SMOOTH);
identify var = logprice crosscor = time noprint;
estimate input = time p = (1);
run;

Residual Plots




Residuals for logprice
0.10
0.05
0.00
-0.05
-0.10
Residual
0
20
40
60
80
100
Observation
Loess Smoother