Predicting the Stock Market: How Regression can be Applied to Achieve a Parsimonious Model

Nader Liddawi
nliddawi@depaul.edu

***Abstract*—**I want to find a mathematical model that can predict the stock market using a number of related but distinct variables. The financial market is made of a diverse but, oftentimes, interrelated set of products that move together, opposite of one another or totally randomly of one another. I collect my data sets, carry out auto selection on them and diagnose my results to avoid issues of multi-collinearity, non-constancy, non-normality, outliers, over-fitting, dependence, etc.

1 INTRODUCTION

My motivation for this statistical study is to analyze various types of financial assets with respect to the overall market. Some assets move in lockstep with the overall market, represented by the SPX index, while some assets do not. Regressing SPX on various assets to find a mathematical model can help me predict the value of the SPX at any time given I have the values of the predictor variables.

2 DATA SETS

I need to first collect the data sets of interest. My sources are from the Yahoo Finance and Nasdaq. My 216 data points represent the closing prices of the response and predictor variables since the beginning the of the 2021 year (i.e., year-to-date closing prices).

2.1 Response variable

My response variable will be the stock market as a whole, namely SPX. The SPX index tracks the collection of the largest 500 stocks in the United States, as determined by market capitalization (the product of the number of shares outstanding and the share market price).

Figure 1: SPX index chart

2.2 Predictor Variable

My predictor variables number 7 and are of various asset classes.

2.2.1 *Bitcoin*

The BLX liquidity index is the spot price of the bitcoin cryptocurrency at which a market participant can easily enter and exit (i.e., liquid price).

Figure 2: Bitcoin chart

2.2.2 *EUR/USD*

The EUR/USD is a foreign exchange currency pair representing the price the Euro relative to the U.S. dollar. Quoted in the second currency, namely the USD, the price of 1.16 indicates that a market participants must pay 1.16 USD to buy 1 Euro (making the Euro stronger than the U.S. dollar).

Figure 3: EUR/USD chart

2.2.3 *Ten Year Treasury Yield*

The TNX is an index of the 10-Year Treasury Yield, quoted in the multiple of 10. It represents the market yield on the Treasury Note.

Figure 4: 10-Year Yield chart

2.2.4 *Microsoft*

Microsoft is a large cap software technology stock, or ownership claim on the company assets, that trades on the Nasdaq stock exchange.

Figure 5: Microsoft stock chart

2.2.5 *Tesla*

Tesla is another large cap and automotive technology stock, or ownership claim on the company assets, that trades on the Nasdaq stock exchange.

Figure 6: Tesla stock chart

2.2.6 *ES Futures*

The /ES futures is a futures contract that obligates the owner of the contract to buy the financial asset at his price of entry and the seller of the contract to sell the financial asset at his price of entry. In the derivatives market, these contracts closely follow the underlying asset, namely the SPX index, since the futures contract and the underlying asset converge in price at the time of the December expiration. Cash-settled, the holders of the contract realize a gain or loss at expiration based on where the price of the SPX lies at the December expiration and at what price they entered as long buyers or short sellers.

Figure 7: ES futures chart (Dec expiry)

2.2.7 *Gold Futures*

The /GC futures is a gold contract whose underlying is gold, similar to the previous /ES contract. The long holder will take delivery from the short holder of the underlying gold asset at December expiration and the short holder will deliver the underlying gold asset to the long holder.

Figure 8: Gold futures chart (Dec expiry)

3 ANALYSIS

3.1 ANOVA on All Variables

I regress SPX on the other variables and retrieve the ANOVA table below.

```
*read excel file;
proc import out=set1 datafile= "/home/u59424498/sasuser.v94/Final Project/final proj data.xlsx" DBMS=xlsx REPLACE;
run;

*regress SPX on other variables and output ANOVA with beta parameters;

proc reg data=set1;
title 'ANOVA before Model Selection';
model SPX = BITCOIN MSFT EUR_USD ES_Futures Ten_Year_Yield TSLA Gold_Futures /clb alpha=0.05;
*output out=out_set1 r=residual;
run;

```

ANOVA before Model Selection

The REG Procedure
Model: MODEL1
Dependent Variable: SPX SPX

Number of Observations Read	216
Number of Observations Used	216

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	13893218	1984745	64428.2	<.0001
Error	208	6407.54806	30.80552		
Corrected Total	215	13899626			

Root MSE	5.55027	R-Square	0.9995
Dependent Mean	4206.95597	Adj R-Sq	0.9995
Coeff Var	0.13193		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	166.52656	50.44316	3.30	0.0011	67.08117	265.97194
BITCOIN	BITCOIN	1	-0.00012382	0.00006565	-1.89	0.0607	-0.00025325	0.00000561
MSFT	MSFT	1	-0.04645	0.07985	-0.58	0.5614	-0.20387	0.11096
EUR_USD	EUR_USD	1	-133.85152	48.04947	-2.79	0.0058	-228.57791	-39.12512
ES_Futures	ES_Futures	1	0.99651	0.00827	120.54	<.0001	0.98021	1.01281
Ten_Year_Yield	Ten_Year_Yield	1	7.10792	3.32644	2.14	0.0338	0.55005	13.66579
TSLA	TSLA	1	0.00917	0.00508	1.81	0.0723	-0.00083730	0.01917
Gold_Futures	Gold_Futures	1	0.00840	0.01275	0.66	0.5106	-0.01673	0.03353

Figure 9: ANOVA on all variables

The ANOVA table calculates the adjusted R-Square to be 0.9995, notably high for predictability of the response variables given the other variables.

This figure shows that under the alpha=5%, the significant predictor variables are EUR_USD, ES_Futures and Ten_Year_Yield. However, as we delete the other seemingly insignificant predictor variables, all the variables' p-values will change. So, we need to detect this p-value change using auto-selection.

3.2 Stepwise Selection

We begin with a table with no entries. The smallest p-value variable gets inserted, as long as it is under our p-value threshold of 0.15. And the table is then examined to see if any variables have a p-value of greater than a threshold of 0.30, in which case we delete from the table. We do this iteratively until no more predictor variables are left to be examined. We can see that this stepwise auto-selection process produced two predictor variables, namely ES_Futures and EUR_USD.

```
19 proc reg data=set1;
20 title 'Stepwise Selection';
21 model SPX = BITCOIN MSFT EUR_USD ES_Futures Ten_Year_Yield TSLA Gold_Futures / details slentry=0.15 slstay=0.30 selection=stepwise;
22 output out=out_set1 r=residual;
23 run;
```

All variables left in the model are significant at the 0.3000 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ES_Futures		ES_Futures	1	0.9995	0.9995	16.6490	422084	<.0001
2	EUR_USD		EUR_USD	2	0.0000	0.9995	6.4046	12.05	0.0006

Figure 10: Stepwise Summary

3.3 Potential Outliers

We need to determine whether there are any potential outliers to remove. Outliers, or data points that deviate from the normal trend, can cause us to mis-interpret the parameter estimates and cause our residual diagnostics to look like they violate our residual assumptions.

For example, the below output shows that normality and constancy assumptions are not met.

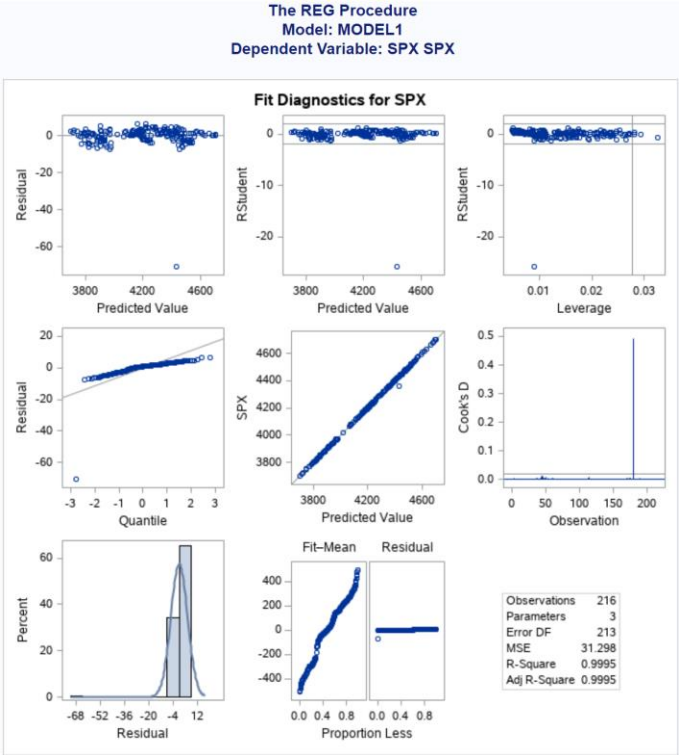


Figure 11: Residual Diagnostic shows outlier

Notice above that there exists a data point in the constancy chart towards the end of the data collection that seems like an outlier.

3.4 Influential Diagnostics

We can find the outliers by examining three different estimates of the influence of a data point in our data collection. We want to compare the critical value of the estimate with the calculated value of the estimate, for each of our 216 data points. The three right-most columns below represent the estimates: Studentized Residuals, DFFITS, and Cook's D.

```
27 title 'Examining Potential Outliers';
28 ods output OutputStatistics=residual1;
29
30 proc reg data=set1 plots=(DFFITS DFBETAS);
31 model SPX = EUR_USD ES_Futures;
32
33 output out=out_set1 dffits=DFFITS cookd=CookD rstudent=R_Student predicted=Predicted;
34 run;
35
36
37 proc print data=out_set1;
38 run;
```

Examining Potential Outliers

Obs	Observation	SPX	BITCOIN	MSFT	EUR_USD	ES_Futures	Ten_Year_Yield	TSLA	Gold_Futures	Predicted	CookD	R_Student	DFFITS
1	1	3700.65	31971.91	216.28	1.22507	3692.2	0.917	729.77002	1946.6	3698.18	0.00162	0.4463	0.06965
2	2	3726.86	33992.43	216.48	1.22516	3718.2	0.955	735.109985	1954.4	3724.03	0.00200	0.5114	0.07728
3	3	3748.14	36824.36	210.87	1.230027	3740.6	1.042	755.97998	1908.6	3745.85	0.00138	0.4135	0.06411
4	4	3803.79	39371.04	216.87	1.234111	3795.6	1.071	816.039978	1913.6	3800.17	0.00374	0.6554	0.10573
5	5	3824.68	40797.61	218.19	1.227144	3817.5	1.105	880.02002	1835.4	3822.59	0.00095	0.3763	0.05332
6	6	3799.61	35566.66	216.08	1.218621	3792	1.132	811.190002	1850.8	3798.03	0.00047	0.2851	0.03751
7	7	3801.19	33922.96	213.53	1.21607	3794.5	1.138	849.440002	1844.2	3800.75	0.00004	0.0794	0.01030
8	8	3809.84	37316.36	214.93	1.220889	3803.8	1.088	854.409973	1854.9	3809.55	0.00002	0.0523	0.00690
9	9	3795.54	39187.33	211.64	1.216249	3791.2	1.129	845	1851.4	3797.45	0.00067	-0.3437	-0.04486
10	10	3768.25	36825.37	211.27	1.215126	3762.25	1.097	826.159973	1829.9	3768.77	0.00005	-0.0927	-0.01273

Figure 12: Influential Diagnostics for first 10 observations

3.4.1 *Influential Diagnostics: Studentized Residuals*

Using a sample size of 216, an alpha of 5%, the number of parameter estimates at 3 (which includes the B0, B1 and B2), we can find the studentized residual critical value of +/- 3.74611.

```
57 title 'Observations that have possible outlying Y observations';
58 proc print data=resid1;
59 where Rstudent>3.74611 or Rstudent<-3.74611;
60 var Observation Rstudent;
61 run;
```

T distribution values

Obs	alpha	n	p	t_stud_del_res2
1	0.05	216	3	3.74611

Figure 13: Studentized Residual critical value

As we go down the table of all of our 216 observations, we find but one observation that exceeds this critical value, namely observation 180.

Obs	Observation	SPX	BITCOIN	MSFT	EUR_USD	ES_Futures	Ten_Year_Yield	TSLA	Gold_Futures	Predicted	CookD	R_Student	DFFITS
180	180	4357.73	42843.8	294.3	1.172993	4421.75	1.309	730.169983	1763.8	4428.53	0.49129	-25.8227	-2.46597

Figure 14: Observation 180 influential output

Observation 180 has a calculated value of -25.8227, which of course exceeds the critical value of 3.74611. So, this is a candidate for deletion due to its outlier nature.

3.4.2 *Influential Diagnostics: DFFITS*

Another measure we can use to detect an outlier is the DFFITS.

```
66 title 'Observations that have a large DFFITS values';
67 * 2sqrt(3/216) = 0.2357;
68 proc print data=out_set1;
69 where DFFITS>0.2357 or DFFITS<-0.2357;
70 var Observation DFFITS;
71 run;
```

Our DFFITS critical value is determined by 2*sqrt(p/n), which equals +/- 0.2357 in our case. Again, only one observation happens to exceed these critical bounds, namely observation 180.

Observations that have a large DFFITS values

Obs	Observation	DFFITS
180	180	-2.46597

Figure 15: DFFITS critical value

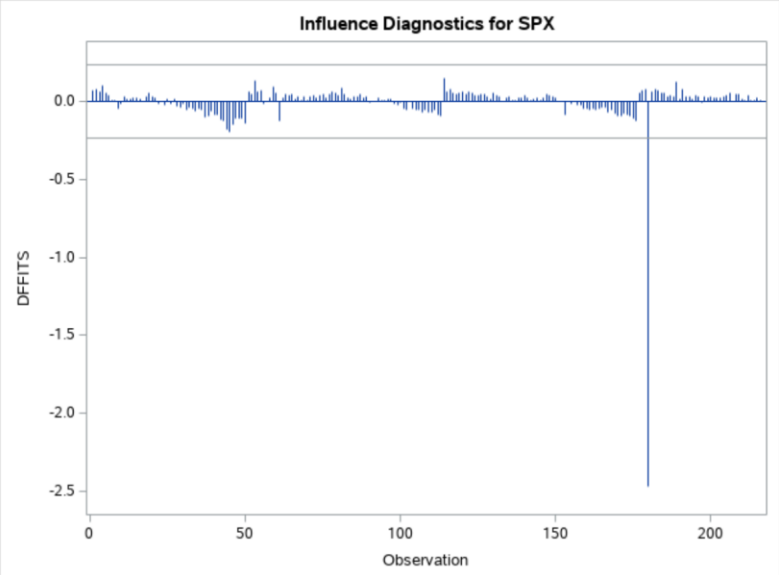


Figure 16: DFFITS chart showing potential outlier

3.4.3 *Influential Diagnostics: Cook's D*

Our final influential diagnostic measure is Cook's D. Interestingly, we find that no observation exceeds Cook's D critical value of 0.7912.

```
75 title 'Observations that have a large COOK''s Distance values';
76 proc print data=out_set1;
77 where COOKD>0.7912;
78 var Observation CookD;
79 run;
```

In fact, observation 180 is at 0.49129, which is not greater than 0.7912.

Obs	Observation	SPX	BITCOIN	MSFT	EUR_USD	ES_Futures	Ten_Year_Yield	TSLA	Gold_Futures	Predicted	CookD	R_Student	DFFITS
180	180	4357.73	42843.8	294.3	1.172993	4421.75	1.309	730.169983	1763.8	4428.53	0.49129	-25.8227	-2.46597

Figure 17: Observation 180 influential output

3.5 Residual Diagnostics

After we delete the observation 180, we find that our residual assumptions of normality, constancy and independence are satisfied.

```
86 data set2;
87 set set1;
88 if _n_=180 then delete;
89 run;
90
91 title 'PROC REG output WITHOUT POTENTIALLY EXTREME OBSERVATION 180';
92 proc reg data=set2;
93 model SPX = EUR_USD ES_Futures /r partial influence vif clb alpha=0.05;
94 output out=out_set2 r=residual;
95 run;
```

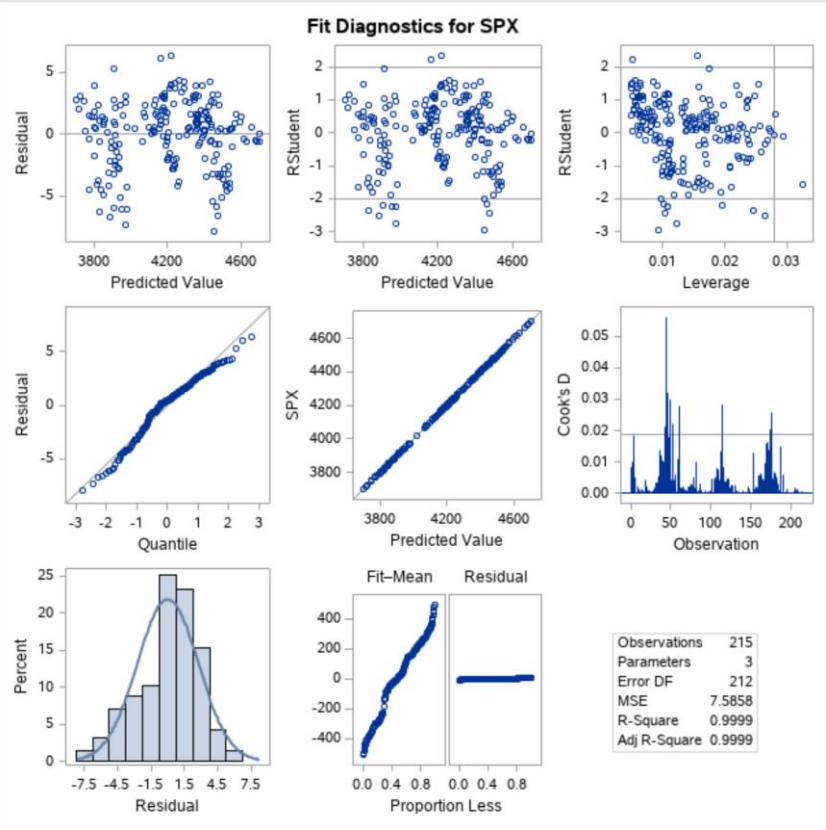


Figure 18: Residual Diagnostics after deletion of observation 180

Notice that the constancy chart (first row, first column) is scattered about without a noticeable pattern. The outlier chart (first row, second column) is also scattered about with no pattern and most of the data points falling within two standard deviations and about 5% falling outside two standard deviations. This pattern is in keeping with the empirical rule. Moreover, the quantile chart (second row, first column) shows the data points hugging the line, which indicates normality. And the histogram (third row, first column) indicates more-or-less normality. Perhaps with a greater number of data points, we can see a slight left-skewedness diminish in favor of a more normal curve.

```
*output Time Series (residual vs. observations) to check any seasonal patterns;
proc gplot data=out_set3;
plot residual*order / vaxis=axis1 haxis=axis2;
title "Sequence plot of the residuals";
axis1 label = (a=90 'Residual');
axis2 label=('Observation number');
symbol v=dot cv=blue ci=red i=join;
run;
```

Now, notice below the lack of a time-series pattern when we plot the observations in order against their residuals.

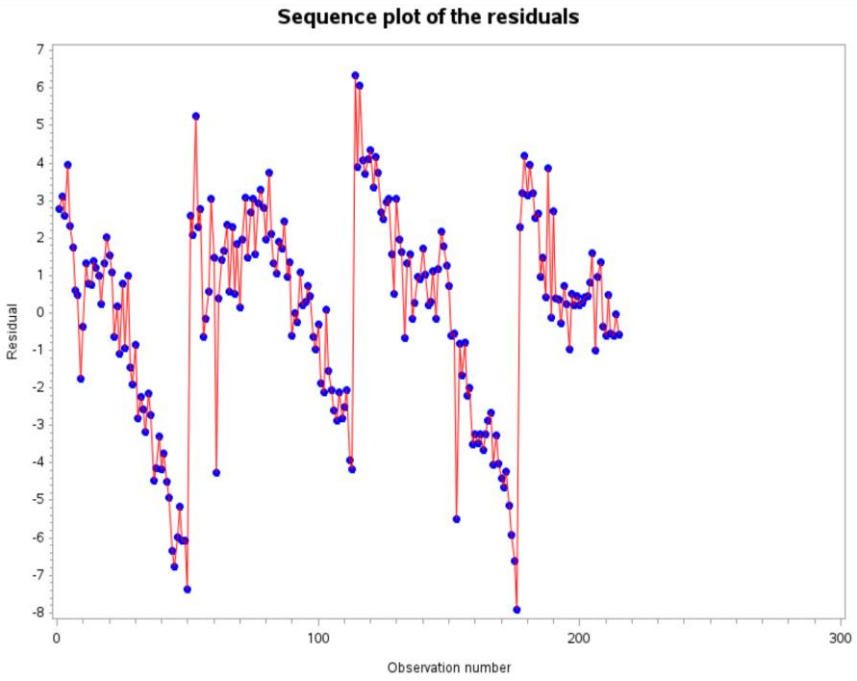


Figure 19: Independence Chart

3.6 Multi-Collinearity

Multi-collinearity occurs when an X value can be linearly predicted by another X value. This can mess with our multiple regression analysis by having some coefficient estimates depend on each other, so finding Variance Inflation Factors above 10 (indicating multi-collinearity) is critical. Deletion of one X term and keeping another one may be in order.

```
91 title 'PROC REG output WITHOUT POTENTIALLY EXTREME OBSERVATION 180';
92 proc reg data=set2;
93 model SPX = EUR_USD ES_Futures /r partial influence vif clb alpha=0.05;
94 run;
```

From our table below, we see that no VIF is above 10 and therefore we do not have multi-collinearity. In fact, EUR_USD and ES_Futures are both under 2.

PROC REG output WITHOUT POTENTIALLY EXTREME OBSERVATION 180

The REG Procedure
Model: MODEL1
Dependent Variable: SPX SPX

Number of Observations Read	215
Number of Observations Used	215

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	13875179	6937589	914552	<.0001
Error	212	1608.18502	7.58578		
Corrected Total	214	13876787			

Root MSE	2.75423	R-Square	0.9999
Dependent Mean	4206.25470	Adj R-Sq	0.9999
Coeff Var	0.06548		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	150.48161	19.08062	7.89	<.0001	0	112.86956	188.09366
EUR_USD	EUR_USD	1	-103.19798	13.19193	-7.82	<.0001	1.99740	-129.20213	-77.19383
ES_Futures	ES_Futures	1	0.99502	0.00105	951.40	<.0001	1.99740	0.99296	0.99708

Figure 20: ANOVA on the two predictor variables found from Stepwise Selection

3.7 Regression Model

After completing our influential and residual diagnostics, we can now formulate the regression model. Following the familiar Y=m*x + b format, we can transform the above data into the following regression model: Y-hat = 150.48 -103.20 * EUR_USD + 0.995 * ES_Futures.

Notice how after we deleted the observation 180 and carried out stepwise selection our adjusted R-Squared increased slightly. It was previously 0.9995, and now it is 0.9999.

4 RESULTS

On the whole, we came up with a sound regression model that can accurately predict where the SPX will be if we know the values of the predictor variables. Deleting the extreme observation and insignificant predictor variables helped increase our adjusted R-Squared measure.

4.1 Future Case

I suppose that in future cases, I would not select a predictor variable that, though distinct, closely follows the response variable. We see below that the correlation in the added-variable plots shows how ES_Futures and SPX are closely correlated due to their nature.

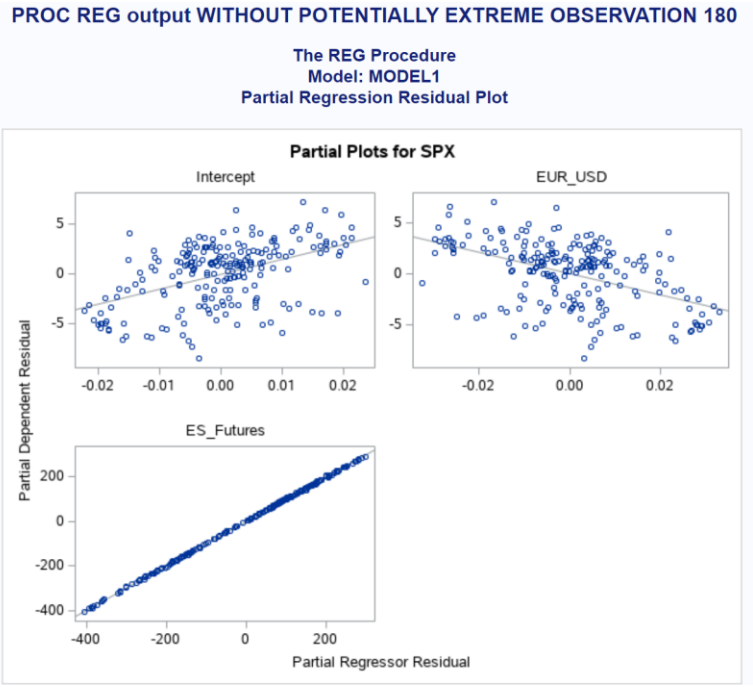


Figure 21: Partial Regression Residual Plots

4.2 Stepwise over All

I chose to select stepwise selection over the backward selection because the former resulted in 2 predictor variables while the latter resulted in 5 predictor variables. (Forward selection gave me the same model as the stepwise selection.) Parsimony, or having the least number of predictor variables while keeping predictability high, was a key attribute in my decision.

Moreover, I chose to select stepwise over the AIC/BIC method which penalizes for the insertion of predictor variables because of parsimony.

```
13 proc reg data=set1;
14 title 'AIC/BIC Selection';
15 model SPX = BITCOIN MSFT EUR_USD ES_Futures Ten_Year_Yield TSLA Gold_Futures / selection=rsquare adjrsq cp aic bic;
16 run;
```

Notice below, the chart with the lowest AIC/BIC measures and the highest adjusted R-Squared measure includes 5 predictor variables.

AIC/BIC Selection

The REG Procedure
Model: MODEL1
Dependent Variable: SPX

R-Square Selection Method

Number of Observations Read	216
Number of Observations Used	216

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	BIC	Variables in Model
5	0.9995	0.9995	4.8594	745.1206	747.5285	BITCOIN EUR_USD ES_Futures Ten_Year_Yield TSLA

Figure 22: AIC/BIC Selection output

Had I chosen these variables instead of the 2 I did choose (namely, EUR_USD and ES_Futures), I would not have kept parsimony at the heart of my study.

4.3 ES_Futures Deleted

If I had not regressed SPX on ES_Futures (i.e., deleted ES_Futures), I would have ended up with all the other 6 predictor variables in my stepwise selection summary, per below.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	MSFT		MSFT	1	0.9123	0.9123	349.024	2226.85	<.0001
2	TSLA		TSLA	2	0.0334	0.9457	137.244	131.13	<.0001
3	EUR_USD		EUR_USD	3	0.0081	0.9538	87.3406	37.26	<.0001
4	Ten_Year_Yield		Ten_Year_Yield	4	0.0088	0.9626	33.0534	49.68	<.0001
5	Gold_Futures		Gold_Futures	5	0.0042	0.9669	7.9440	26.86	<.0001
6	BITCOIN		BITCOIN	6	0.0005	0.9673	7.0000	2.94	0.0877

Figure 23: Stepwise Summary in scenario where ES_Futures is not analyzed or regressed

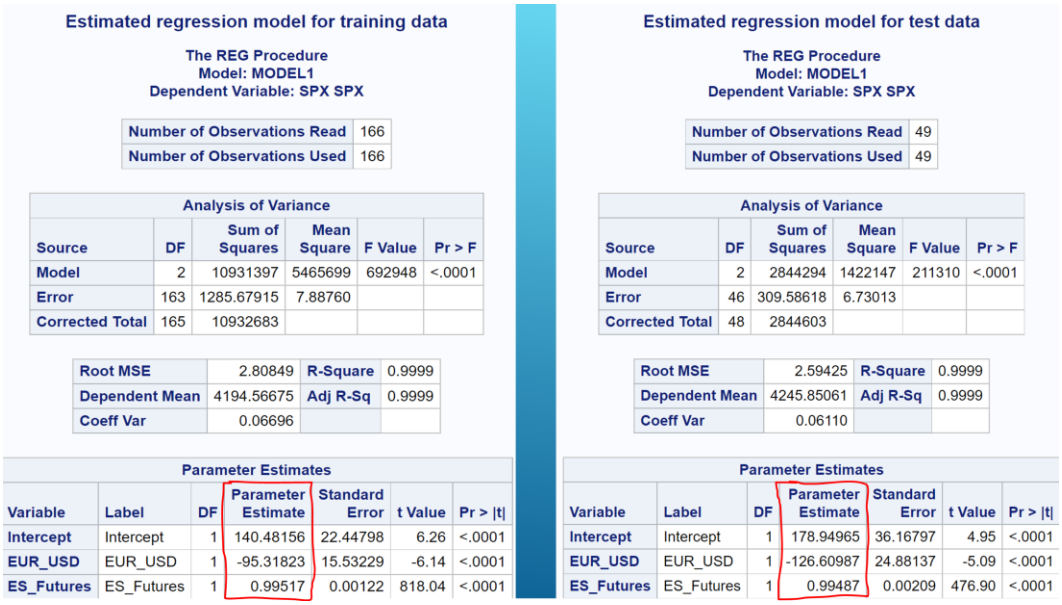
I would probably have had to change my 0.15 and 0.30 threshold to some number lower to arrive at a model with maybe 3 or 4 predictor variables. Otherwise, without including ES_Futures, all the other variables would be in my model and parsimony would not have been achieved.

4.4 Model Validation

Nonetheless, I am satisfied that my chose regression model does hold up to new or test data. Over-fitting, or forming a regression model that predicts a current sample very well but fails to predict other samples from the same population as well, can be a problem.

So, we conducted model validation by randomly splitting our data set into two. The first data set will consist of about 80% of the original data, called the training

data set. And the second data set will consist of about 20% of our original data, called the test data set. We regress separately the SPX on these two data sets using our original 2 predictor variables and output the ANOVA tables side-by-side.



Estimated regression model for training data

The REG Procedure
Model: MODEL1
Dependent Variable: SPX SPX

Number of Observations Read	166
Number of Observations Used	166

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	10931397	5465699	692948	<.0001
Error	163	1285.67915	7.88760		
Corrected Total	165	10932683			

Root MSE	2.80849	R-Square	0.9999
Dependent Mean	4194.56675	Adj R-Sq	0.9999
Coeff Var	0.06696		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	140.48156	22.44798	6.26	<.0001
EUR_USD	EUR_USD	1	-95.31823	15.53229	-6.14	<.0001
ES_Futures	ES_Futures	1	0.99517	0.00122	818.04	<.0001

Estimated regression model for test data

The REG Procedure
Model: MODEL1
Dependent Variable: SPX SPX

Number of Observations Read	49
Number of Observations Used	49

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2844294	1422147	211310	<.0001
Error	46	309.58618	6.73013		
Corrected Total	48	2844603			

Root MSE	2.59425	R-Square	0.9999
Dependent Mean	4245.85061	Adj R-Sq	0.9999
Coeff Var	0.06110		

Parameter Estimates

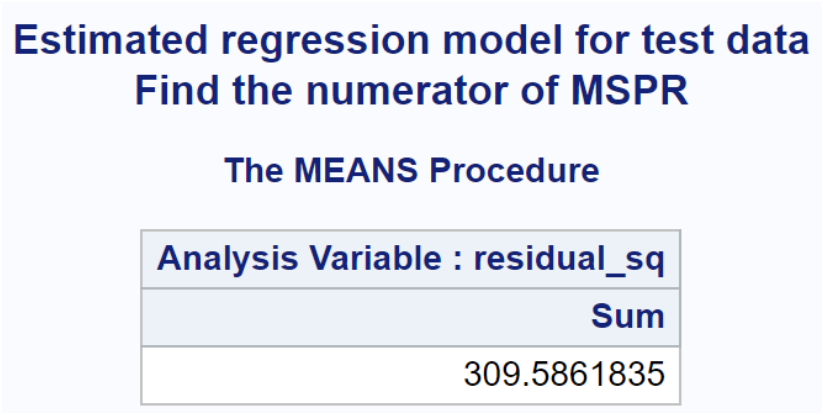
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	178.94965	36.16797	4.95	<.0001
EUR_USD	EUR_USD	1	-126.60987	24.88137	-5.09	<.0001
ES_Futures	ES_Futures	1	0.99487	0.00209	476.90	<.0001

Figure 24: ANOVA on Training and Test data sets for Model Validation

Notice above how the parameter estimates are similar and have the same signs as each other, while holding adjusted R-Squared high at 0.9999. This indicates that the original parameter estimates can predict very well new sample data sets from the same population.

Code for the above tables and the MSPR can be found in the Appendix below.

5 APPENDIX



Estimated regression model for test data
Find the numerator of MSPR

The MEANS Procedure

Analysis Variable : residual_sq
Sum
309.5861835

Figure 25: MSPR output (numerator only)

- MSPR = 309.5862 / 49 = 6.3181
- MSE from Training data set was 7.8876
- Since MSPR and MSE(Training) are close, our model has good predictive power and does not pose the problem of over-fitting

```
*MODEL VALIDATION;

data set_new;
set set2;
* Generate a uniform random number between 0 and 1;
unif = ranuni(720034930);
if unif <=0.80 then build = 1;
else build = 0;
run;

data training;
set set_new;
where build=1;
run;

data test;
set set_new;
where build=0;
run;

title 'Estimated regression model for training data';
proc reg data=training;
model SPX = EUR_USD ES_Futures;
run;


title 'Estimated regression model for test data';
proc reg data=test;
model SPX = EUR_USD ES_Futures;
output out=out_set4 predicted=predict;
run;

proc print data=out_set4;
var SPX predict;
run;


data out_set5;
set out_set4;
residual_sq = (SPX - predict)**2;
run;
```

```
title2 'Find the numerator of MSPR';
proc means data=out_set5 sum;
var residual_sq;
run;

quit;
```