Multiple Regression:

The Noble venture of knowing the past and predicting the future

Nomuntuya Luehr

Statistics UB-103

Prof. Lawrence Tatum

Table of Contents

Project Assignment 1

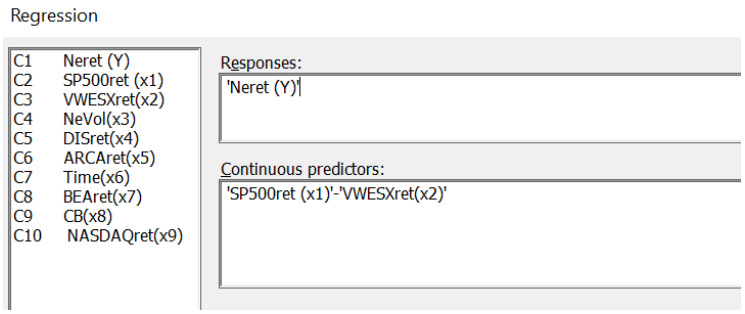
Mostly likely, you have been a customer of the Noble Corporation. As a part of the petroleum industry, 49% of this offshore drilling company's revenues come from its business with Shell Oil Company, which has over 14,000 locations in the U.S. Even though it has only 3,300 employees, it collects $3.35 billion in revenues each year. Since its inception in 1985, it has only had one accident, in which one drillship lost its mooring and drifted close to shore. Thankfully, the accident didn't result in any injuries or environmental damage.

Before running a regression model, we must first calculate our y-value by finding the monthly percentage returns of the Noble Corporation (Neret%). The first ten rows of the monthly percentage returns of my stock is shown below.

Date	Open	High	Low	Close	Volume	Adj Close	Neret
11/1/2016	5.03	6.45	4.45	6.22	13435100	6.22	-20.5788
10/3/2016	6.33	6.51	4.82	4.94	13177800	4.94	28.34008
9/1/2016	5.78	6.7	5.09	6.34	12559000	6.34	-9.14826
8/1/2016	7.27	7.51	5.59	5.76	9052200	5.76	28.125
7/1/2016	8.24	8.98	6.99	7.38	8418700	7.38	11.34888
6/1/2016	8.23	9.73	7.82	8.24	9406900	8.217547	1.21361
5/2/2016	11.25	11.35	8.01	8.34	10387500	8.317276	34.65226
4/1/2016	10.04	12.19	9.18	11.23	11591900	11.1994	-7.99001
3/1/2016	8.23	13.9	8	10.35	13907600	10.30457	-19.5169
2/1/2016	7.42	9.17	6.66	8.33	9811600	8.293434	-8.3381

The regression equation of the multiple regression model of Neret%, the y-variable, and the X1 variable of the S&P monthly return (SP500%) and the X2 variable of the VWESX monthly return (VWESX%) is calculated as followed through Minitab.

Step1: Input Data



Step2: Output Data

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
11.2281	26.25%	24.99%	21.83%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	
Constant	2.11	1.04	2.03	0.044	
VWESXret	-0.549	0.367	-1.49	0.138	1.04

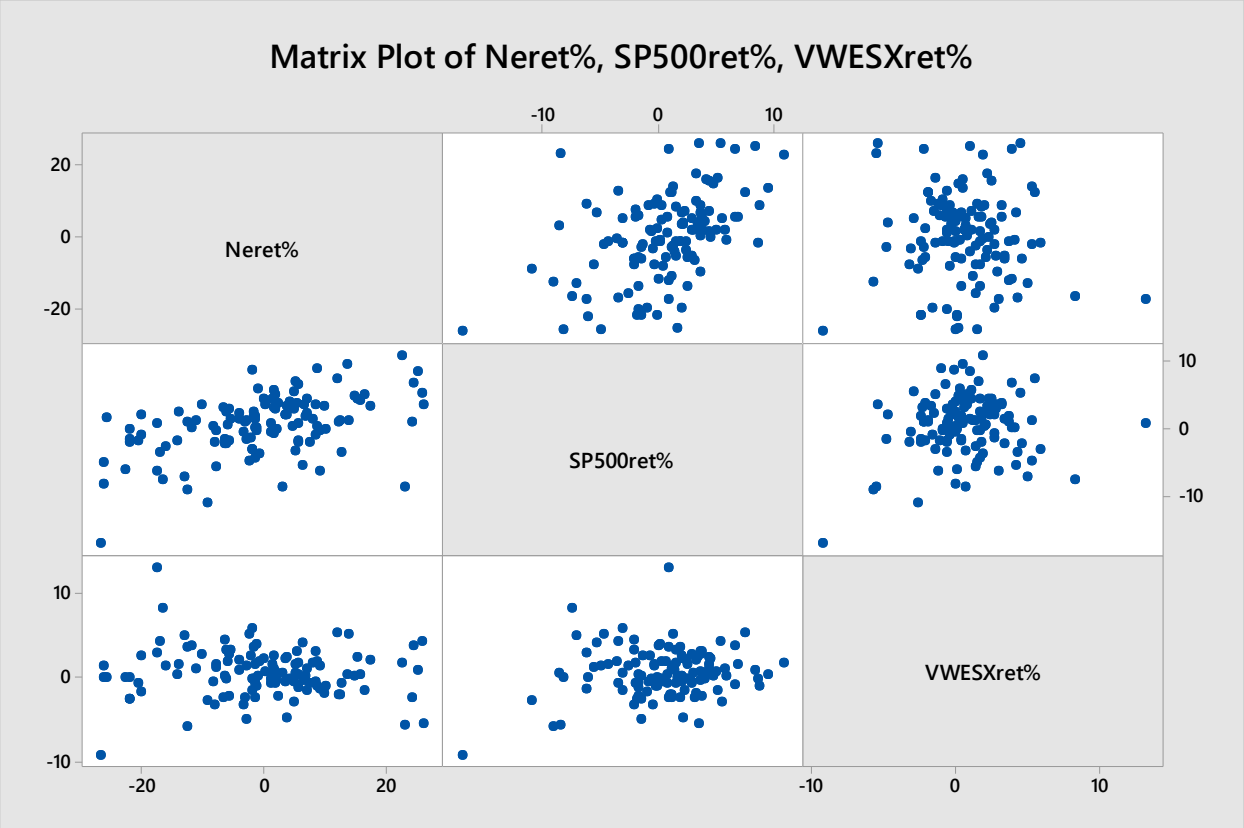
```
S&P500ret    1.481     0.230      6.45     0.000  1.04
```

Regression Equation

Neret = 2.11 – 0.549 VWESXret + 1.481 S&P500ret

The regression software models the relationship between E(Y) and x-variable values through the equation $E(Y) = A + B_1 * (X_1) + B_2 * (X_2)$. The y-intercept plus the slope of the Neret is against the S&P500ret% * S&P500ret% plus the slope of the Neret% against the VWESXret% * VWESXret%.

Using Minitab, I created a matrix scatterplot (shown below) for Y, X1, and X2. The horizontal variables are X1, the S&P500ret%, and X2, the WVESXret%. The vertical variable is Y, the Neret%.



From the regression model calculated above, I found the Y- hat equation, which is the same equation as my regression equation. The Y-hat represents the predicted equation for a line of best fit in the regression and is represented by the form y-hat = a +bx, where a is the y-intercept and b is the slope. This equation differentiate the predicted/fitted data from the observed data y, which in this case is the daily return percentage of my stock (Neret%)

Output:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
11.2281	26.25%	24.99%	21.83%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	
Constant	**2.11**	1.04	2.03	0.044	
VWESXret	**-0.549**	0.367	-1.49	0.138	1.04
S&P500ret	**1.481**	0.230	6.45	0.000	1.04

Equation: Y-hat = 2.11 – 0.549x1 **+ 1.481**x2

Project Assignment 2

In this phase, we will continue to explore the nature of multiple regression. However, this time we will use the following nine x-variables versus 2, and the same y-variable as in section 1.

Variables

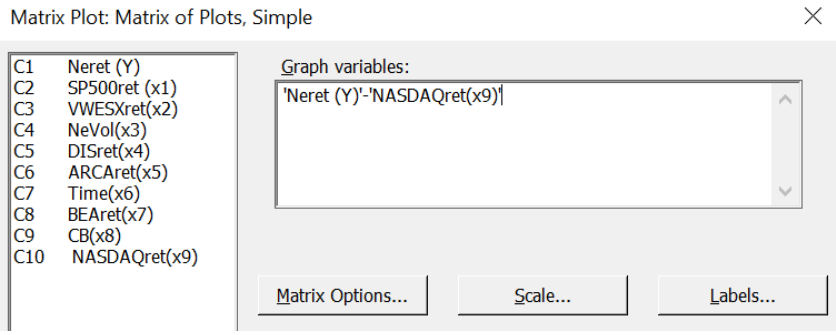
- Y: Monthly percentage return of Noble Corporation NYSE equity stock
- X1: Monthly % return on the S&P500
- X2: Monthly % return on the VWESX
- X3: Monthly change in the trading volume of Noble Corporation NYSE equity stock
- X4: Monthly % return on Disney NYSE equity
- X5: Monthly % return on ARCA Computer Tech Index NYSE non-equity security
- X6: Time in months (1-120)
- X7: Monthly % change in a BEA econometric measure
- X8: Monthly % change in a Census Bureau econometric measure
- X9: Monthly % return on the NASDAQ

Below are the first 10 rows of my enlarged data set:

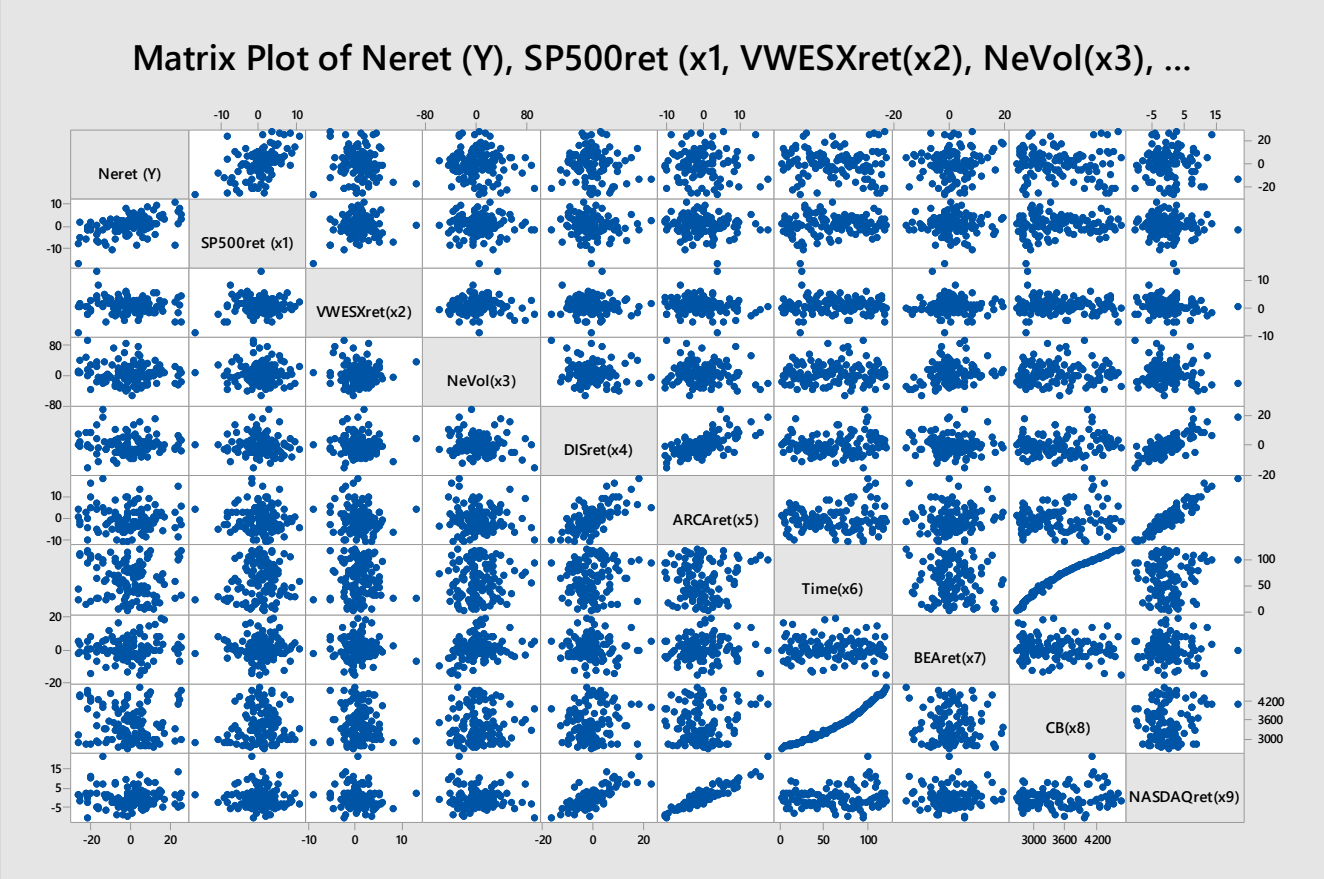
	Neret (Y)	SP500ret (x1)	VWESXret(x2)	NeVol(x3)	DISret(x4)	ARCAret(x5)	Time(x6)	BEAret(x7)	CB(x8)	NASDAQret(x9)
1	-1.5758	1.4059	-0.5890	-21.1897	-0.8403	1.0219	1	-0.5857	2661.7	-0.1748
2	-6.2593	-2.1846	3.2782	-1.9151	-6.4871	0.2289	2	16.1658	2690.6	-2.5272
3	12.0478	0.9980	-1.9647	-4.6958	0.1834	0.0698	3	-7.5127	2708.2	2.3676
4	7.0285	4.3291	1.0189	-27.9226	1.7230	-2.2410	4	8.4761	2731.0	-1.8596
5	9.7645	3.2549	-1.7857	-6.9983	1.5774	-1.9129	5	1.0821	2727.4	-0.9800
6	5.5526	-1.7816	-1.1790	11.7382	1.2137	-7.1975	6	-8.1257	2749.4	-6.1885
7	5.0656	-3.1982	0.6135	10.4243	1.4312	2.6199	7	9.6230	2766.9	2.1761
8	-4.2358	1.2864	0.9478	11.5947	4.0718	-5.2007	8	-6.1144	2759.8	-3.4901
9	0.0193	3.5794	0.7092	19.9769	-3.8253	6.7086	9	1.0086	2779.3	1.9787
10	7.9511	1.4822	1.7207	-29.4515	-3.8163	-8.5325	10	-1.7782	2764.7	-6.4047
11	-1.4752	-4.4043	1.5802	0.9937	0.3141	2.3284	11	5.5396	2767.6	1.2286

Using Minitab, I created a matrix scatterplot for my Y and X1-X9

Input:



Output:



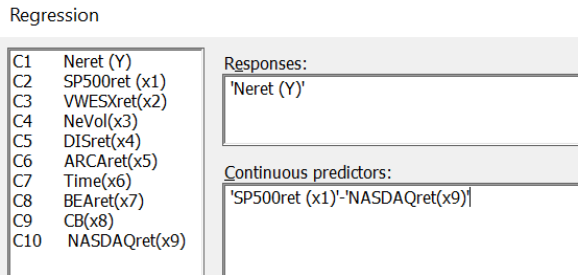
The shapes of the scatterplots appear to be circular and oval. However, for the Time versus Census Bureau, the scatterplot looks similar to a linear regression model. By looking at the distribution, there does not appear to be obvious leverage points or outliers that have skewed the shape of the scatterplots.

In terms of multicollinearity, there is evidence of such a relantionship between DISret (x4), ARCAret (x5), Time (x6), CB (x8), and NASDAQret (x9). This is because the VIF, the

variance inflation factor, which signifies the severity of multi-collinearity is above 1.5 for each of these x values.

Using Minitab, I ran a full regression model for the Y and X1-X9 variables

Input:



Output:

Regression Analysis: Neret (Y) versus SP500ret (x1, VWESXret(x2), NeVol(x3), DISret(x4), ...

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	9	5322.8	591.42	5.48	0.000
SP500ret (x1)	1	4588.0	4587.95	42.53	0.000
VWESXret(x2)	1	326.8	326.84	3.03	0.085
NeVol(x3)	1	176.3	176.32	1.63	0.204
DISret(x4)	1	1.4	1.38	0.01	0.910
ARCAret(x5)	1	58.7	58.68	0.54	0.462
Time(x6)	1	37.8	37.85	0.35	0.555
BEAret(x7)	1	184.5	184.50	1.71	0.194
CB(x8)	1	5.2	5.18	0.05	0.827
NASDAQret(x9)	1	44.3	44.27	0.41	0.523
Error	110	11866.8	107.88		
Total	119	17189.6			

Model Summary

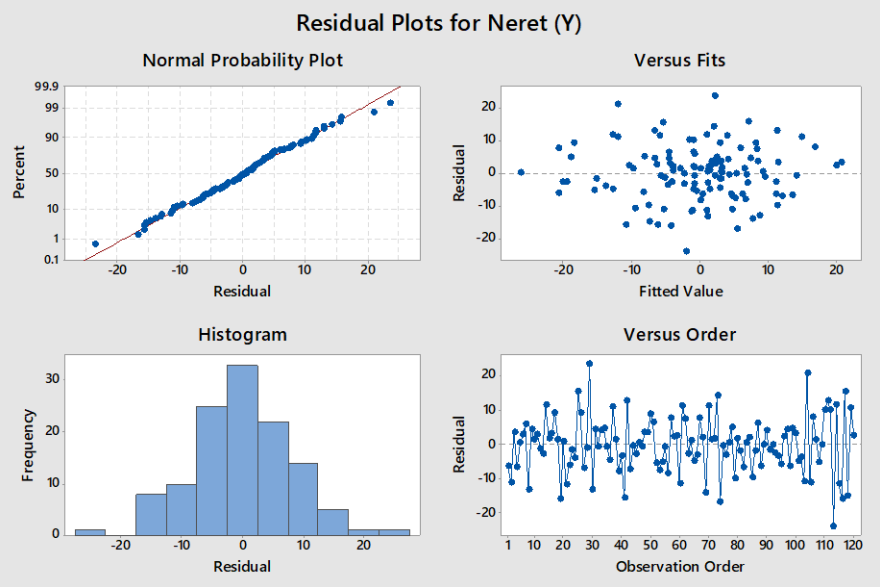
S	R-sq	R-sq(adj)	R-sq(pred)
10.3866	30.96%	25.32%	16.80%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2.3	23.1	-0.10	0.921	
SP500ret (x1)	1.443	0.221	6.52	0.000	1.05
VWESXret(x2)	-0.592	0.340	-1.74	0.085	1.06
NeVol(x3)	-0.0467	0.0365	-1.28	0.204	1.11
DISret(x4)	-0.026	0.233	-0.11	0.910	2.61
ARCAret(x5)	-0.372	0.504	-0.74	0.462	8.51
Time(x6)	-0.089	0.150	-0.59	0.555	29.85
BEAret(x7)	0.196	0.150	1.31	0.194	1.11
CB(x8)	0.00203	0.00927	0.22	0.827	29.89
NASDAQret(x9)	0.391	0.611	0.64	0.523	11.35

```
Regression Equation

Neret (Y) = -2.3 + 1.443 SP500ret (x1) - 0.592 VWESXret(x2) - 0.0467 NeVol(x3)
            - 0.026 DISret(x4) - 0.372 ARCAret(x5) - 0.089 Time(x6) + 0.196 BEAret(x7)
            + 0.00203 CB(x8) + 0.391 NASDAQret(x9)


Fits and Diagnostics for Unusual Observations

Obs  Neret (Y)     Fit    Resid  Std Resid
 19     -20.15    0.66  -20.81       -2.07  R
 25      22.91   -6.90   29.81        3.02  R
100      24.08    1.34   22.74        2.32  R
113     -25.73   -0.24  -25.50       -2.52  R
119      25.91    4.68   21.23        2.21  R

R  Large residual
```

I also ran 4-in-1 diagnostic plots for my stock (Neret%). This includes a Normal Probability Plot, used to detect non-normality, a Histogram, used to detect peaks, outliers, and non-normality, a Versus Fits plot, used to detect non-constant variance, and a Versus Order plot, used to detect the time dependence of residuals.



Using the values occurred from the full regression model, I found the y-hat equation:

```
Y-hat= -2.3 + 1.443 SP500ret (x1) - 0.592 VWESXret(x2)
- 0.0467 NeVol(x3)- 0.026 DISret(x4) - 0.372 ARCAret(x5) - 0.089 Time(x6)
+ 0.196 BEAret(x7) + 0.00203 CB(x8) + 0.391 NASDAQret(x9)
```

I then input the x-values for one month and computed the y-hat value for the month of January 2007. I found the y-hat value to be 5.17566:

```
Y-hat (X-values of January 2007) = -2.3 + 1.443 SP500ret (0.997998) - 0.592 VWESXret(-
 1.96469) - 0.0467 NeVol(-4.69578)- 0.026 DISret(0.183406) - 0.372 ARCAret(0.0698130)
- 0.089 Time(3) + 0.196 BEAret(-7.51270) + 0.00203 CB(2708.2)
+ 0.391 NASDAQret(2.36763)= 5.17566
```

The y-hat value computed above relates to the bell curved distribution of Y given the values of X variables through the residuals. By examining the 4-in-1 graphs (above), I found that the x-axis of the histogram are the residuals. In order to find the residuals we find the difference between the observed variable Y and the predicted Y-hat value.

For the same month of January 2007, I also computed an approximate 95% Prediction Interval for Y given the 9 X-variables. To do so, I took the y-hat value and added and subtracted the standard error of the regression:

Y-hat (X-values of January 2007) = 5.17566

= - Or + 2(Standard error)

= - Or + 2(10.3866) = - 15.5975 and 25.94886

95% Prediction Interval:

[-15.5975, 25.94886]

To verify my answer, I used Minitab to calculate the 95% Prediction Interval:

Fit	SE Fit	95% CI	95% PI
5.17371	3.17309	(-1.11461, 11.4620)	(-16.3491, 26.6965)

My 95% Prediction Interval was close to the exact interval 95% PI computed by Minitab.

95% Prediction Interval calculate above captures the range for where one can expect to see the next data point for one observation. In this specific case, the 95% prediction interval captures the NE monthly percentage return for the next month within the range of given x values, and 95% of the predictions will be in the range. The prediction interval in this case is [-16.3491, 26.6965], which is wider than the confidence interval

95% Confidence Interval captures the range of values based on sample statistics that is 95% likely to contain the value of the unknown population parameter. In this case, the 95% confidence interval captures the unknown population parameter with 95% confidence. This means that 95% of the time, the population parameter will fall in this interval.

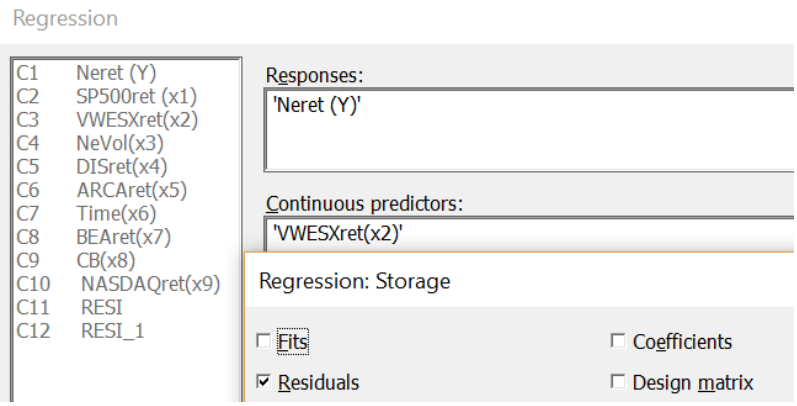
Taking another look at our 4-in-1 plot, the histogram of residual is trying to find the rate of differences between y, our dependent variable, and y-hat, the predicted value of our dependent value generated/assumed by the software.

Project Assignment 3

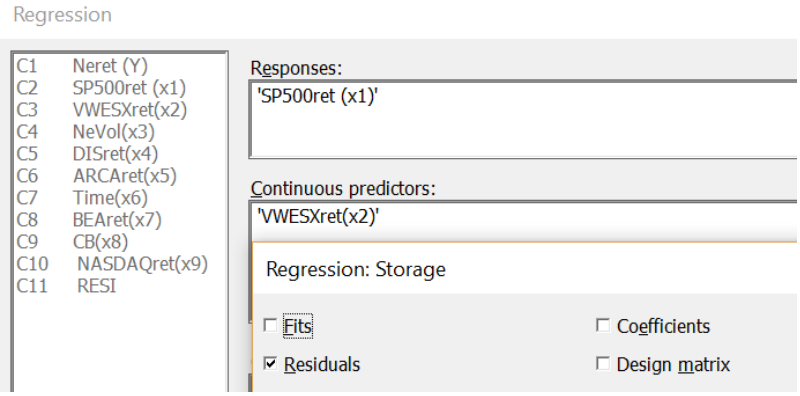
In this section, I found the partial slope of X1 (SP500ret%), using X2(VWESXret%) and my stock as the variable Y (Neret%).

Step 1: SP500ret% is my chosen X1, and VWESXret% is my chosen X2

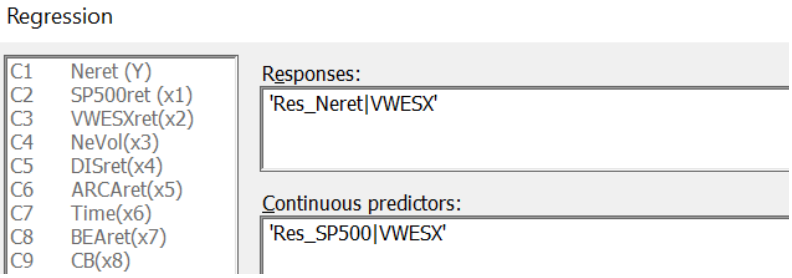
Step 2: Regress Neret% (Y) on VWESXret% (X2), and save the residuals by using the storage function



Step 3: Regress SP500ret% (X1) on VWESXret% (X2), and save the residuals by using the storage function



Step 4: Regress the Res_Neret|VWESX on Res_SP500|VWESX



Step 5: The resulting slope from the regression of 1.396 (highlighted below) is the partial slope of SP500ret% (X1)

Regression Analysis: Res_Neret|VWESX versus Res_SP500|VWESX

```
Analysis of Variance
Source               DF  Adj SS  Adj MS  F-Value  P-Value
Regression            1    4412  4412.4    40.94    0.000
  Res_SP500|VWESX     1    4412  4412.4    40.94    0.000
Error               118   12717   107.8
Total               119   17130
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
```

```
10.3813  25.76%     25.13%      23.08%
Coefficients
Term               Coef  SE Coef  T-Value  P-Value   VIF
Constant         -0.000    0.948    -0.00    1.000
Res_SP500|VWESX   1.396    0.218     6.40    0.000  1.00

Regression Equation
Res_Neret|VWESX = -0.000 + 1.396 Res_SP500|VWESX
Fits and Diagnostics for Unusual Observations

Obs  Res_Neret|VWESX     Fit   Resid  Std Resid
 22           -28.66  -21.21   -7.45      -0.76     X
 25            21.78  -10.67   32.45       3.18  R
 26            -9.72  -14.96    5.25       0.52     X
 29            27.09    5.56   21.53       2.09  R
 58            23.15   14.00    9.14       0.91     X
100            23.76    1.45   22.31       2.16  R
103           -19.20    1.48  -20.68      -2.00  R
113           -25.47    1.65  -27.12      -2.62  R
116           -21.69   -0.65  -21.03      -2.03  R

R  Large residual
X  Unusual X
```

In this multiple regression the casual interpretation of sample slope is that the partial slope of SP500ret% gives the average change in Neret% for a one-unit change in SP500ret%, holding all other variables constant.

However, in rigorous terms, the partial slope for SP500ret% with WVERSret% in the model is the predicted difference between the estimated price of monthly returns on SP500ret% and WVESXret%, which happen to have the same WVESXret% and differ by 1 unit in SP500ret%.

For my y-variable and a subset of x-values x3, x4, x5, and x7, I ran the regression. Then I manually found the R-square value from the given output:

```
                    Analysis of Variance

Source           DF   Adj SS   Adj MS  F-Value  P-Value
Regression        4    470.4  117.591     0.81    0.522
  NeVol(x3)       1    305.0  304.981     2.10    0.150
  DISret(x4)      1      4.4    4.420     0.03    0.862
  ARCAret(x5)     1     13.7   13.702     0.09    0.759
  BEAret(x7)      1    254.2  254.184     1.75    0.189
Error           115  16719.2  145.385
Total           119  17189.6

Model Summary
      S   R-sq  R-sq(adj)  R-sq(pred)
12.0576   2.74%      0.00%       0.00%
```

The full equation I used to find the R-square value is:

R-sq= SSR/SST= SST-SSE / SST 470.4 / 17189.6 = 0.0274 about 2.74%

Step 1: Find SST, the sum of squares for (y – ybar), 470.4
Step 2: Find SSR, the sum of squares from (y – yhat), 17189.6

Step 3: Compute SSR/SST, to get 2.74%

Next, I conducted the F-test using the output given by the multiple regression of Y and X3, X4, X5, and X7 (which is copied below again).

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	**4**	**470.4**	**117.591**	**0.81**	**0.522**
NeVol(x3)	1	305.0	304.981	2.10	0.150
DISret(x4)	1	4.4	4.420	0.03	0.862
ARCAret(x5)	1	13.7	13.702	0.09	0.759
BEAret(x7)	1	254.2	254.184	1.75	0.189
Error	115	16719.2	145.385		
Total	119	17189.6			

Model Summary

S	**R-sq**	R-sq(adj)	R-sq(pred)
12.0576	**2.74%**	0.00%	0.00%

My Null and Alternative Hypotheses are set up as follows:

Ho: B1 = B2 = 0 **Ha**: B does not equal 0

Based on the p-value of the regression (highlighted above), do not reject the null because the p-value of the regression, at 0.522, is greater than 0.05.

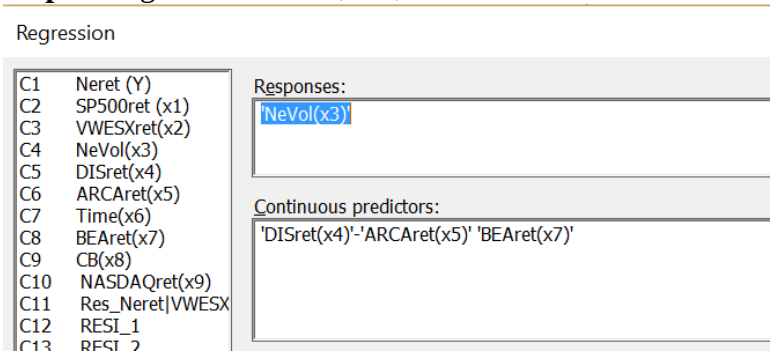
Next, I manually found the VIF, the variance inflation factor, which signifies the severity of multi-collinearity, for my X3 variable:

Manually finding the Variance Inflation Factor of NeVol(x3), the Noble Corporation % change in volume:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.40	1.11	-0.35	0.723	
NeVol(x3)	-0.0606	0.0419	-1.45	0.150	**1.09**
DISret(x4)	-0.040	0.227	-0.17	0.862	1.84
ARCAret(x5)	-0.084	0.273	-0.31	0.759	1.85
BEAret(x7)	0.229	0.173	1.32	0.189	1.10

Step 1: Regress X3 on X4, X5, and X7



Step 2: Find the new R-square value (7.94%)

Regression Analysis: NeVol(x3) versus DISret(x4), ARCAret(x5), BEAret(x7)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	7158.4	2386.1	3.34	0.022
DISret(x4)	1	103.9	103.9	0.15	0.704
ARCAret(x5)	1	230.8	230.8	0.32	0.571
BEAret(x7)	1	6169.8	6169.8	8.63	0.004
Error	116	82974.3	715.3		
Total	119	90132.7			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
26.7450	7.94%	5.56%	0.83%

Step 3: Plug the new R^2 into the equation: $$VIF_i = \frac{1}{1 - R_i^2}$$

1/ (1- R^2x1*x3x4x5) = 1/ 0.9206 = 1.086 or about 1.9

Step 4: Compare with value computed by Minitab:

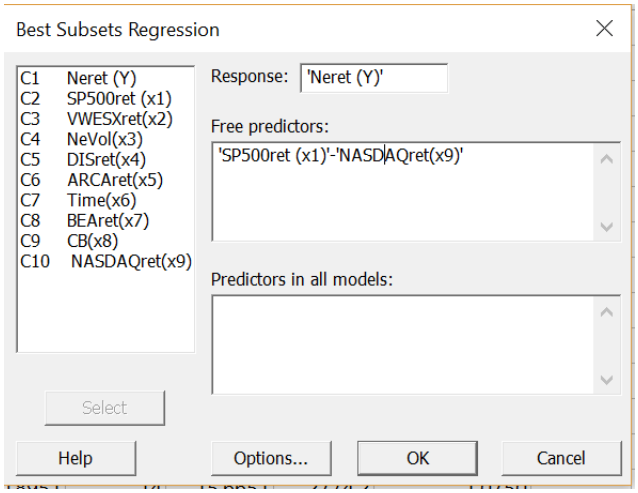
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.40	1.11	-0.35	0.723	
NeVol(x3)	-0.0606	0.0419	-1.45	0.150	**1.09**
DISret(x4)	-0.040	0.227	-0.17	0.862	1.84
ARCAret(x5)	-0.084	0.273	-0.31	0.759	1.85
BEAret(x7)	0.229	0.173	1.32	0.189	1.10

Project Assignment 4

In this section, I explored Mallow's method of model selection that aims to find the best subset of predictors that meet the objective of having a high R-square and a low S value

Input:



Output:

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	500r (x1)	WESXret (x2)	NeVol (x3)	DISSrt (x4)	ARCAre (x5)	ATimet (x6)	BEArt (x7)	CBer (x8)	SDAQret (x9)	
1	24.3	23.6	21.4	4.7	10.505	X									
1	1.2	0.3	0.0	41.5	11.998						X				
2	27.0	25.8	22.5	2.3	10.356	X					X				
2	26.7	**25.4**	**21.8**	**2.8**	10.379	**X**							**X**		**Mallow's Original**
3	28.9	**27.1**	**22.6**	**1.2**	10.262	**X**	**X**				**X**				**Lowest Cp Value**
3	28.7	26.8	22.0	1.7	10.282	X	X						X		
4	29.6	27.2	22.3	2.1	10.256	X	X	X			X				
4	29.4	27.0	22.0	2.4	10.269	X	X				X	X			
5	30.6	27.5	22.3	2.6	10.231	X	X	X			X	X			
5	30.4	27.3	21.8	2.9	10.246	X	X	X				X	X		
6	30.6	26.9	20.9	4.5	10.273	X	X	X			X	X	X		
6	30.6	26.9	20.7	4.6	10.273	X	X	X		X	X	X			
7	30.9	26.6	19.8	6.1	10.296	X	X	X		X	X	X		X	
7	30.7	26.4	19.3	6.4	10.310	X	X	X		X		X	X	X	
8	31.0	26.0	18.3	8.0	10.340	X	X	X		X	X	X	X	X	
8	30.9	26.0	18.3	8.0	10.342	X	X	X	X	X	X	X		X	
9	31.0	25.3	16.8	10.0	10.387	X	X	X	X	X	X	X	X	X	

Acoording to Mallow's original criteria for selecting a model, I picked the one with the Cp value closest to Vars+1, where Vars equals the number of x-regressors. In my specific case, the chosen model includes the SP500% and Census Bureau %, and has a Vars value of 2 and a Cp value of 2.8, which is close to the Vars+1 value of 3. However, according to the practioners preferred method of choosing a model, I chose the one with a Vars value of 3 and a Cp value of 1.2, the lowest of all Cp values. This model includes the SP500%, WVESX%, and Time. The model with the lowest Cp is larger, as it consists of 3 x-values and not 2. When comparing the chosen models, the one chosen according to Mallow has a Cp value that is 1.6 higher than the one with the lowest Cp value. Out of the two competing models, the one with lowest Cp has a larger adjusted R-square value, the proportion of the variance, and a smaller S value, the standard error. Therefore, I prefer the model with the lowest Cp value. This is because a higher adjusted R-square value means that the model fits with my data better. Also, the model has a higher predicted R-square value, which means that the model better predicts responses for new observations. Additionally, this model has a lower S value. This means that it has less standard distance from the fitted regression line and therefore is a better fit.

Extra Credit #1

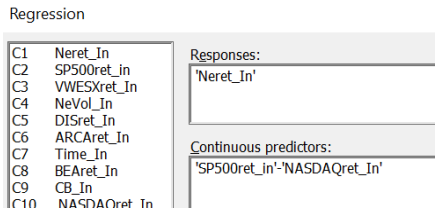
In this section, I look at the "out of sample" data;

Step1: I first begin my setting aside 15% of the rows of my data. The data points that are to be set aside are randomly selected. The first ten rows of my in and out data are shown below, but some x-variables in the "out" section have been cut-off due to limitations.

	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
	VWESXret_In	NeVol_In	DISret_In	ARCAret_In	Time_In	BEAret_In	CB_In	NASDAQret_In	Neret (Y)_1	Neret_out	SP500ret_out	VWESXret_out	NeVol_out	DISret_out	ARCAret_out
1	-2.5283	91.7596	-17.0776	-10.4096	93	4.9587	3928.1	-10.9888	-20.0888	22.9063	-8.56573	-5.56943	-26.5874	-3.0163	2.5557
2	3.5795	-20.7453	-1.6918	-10.6031	76	4.9741	3564.6	-10.7485	8.9540	-7.8933	-0.41886	1.58159	-27.5589	9.7195	2.3891
3	-2.3576	-21.7348	-13.5321	-10.2431	63	-4.9837	3374.0	-10.0212	-16.2058	7.0285	4.32907	1.01888	-27.9226	1.7230	-2.2410
4	-0.0620	-11.0382	-10.1459	-9.9711	15	-4.4212	2778.6	-8.5796	24.9463	24.0819	0.85208	-2.31992	-28.0222	5.4089	14.3655
5	3.6623	-19.3637	-3.5990	-7.2021	60	-11.3402	3319.6	-7.4166	-1.4848	7.9511	1.48223	1.72066	-29.4515	-3.8163	-8.5325
6	0.2785	16.9673	-7.1258	-8.0740	90	-0.5638	3854.8	-7.2509	-20.4974	-5.5715	1.47592	3.28762	-30.4435	-5.6753	-1.6568
7	2.1775	15.0744	-10.5127	-5.7296	82	-2.4192	3677.8	-6.6598	-26.1611	8.4021	-0.86285	-0.49769	-30.7102	9.6639	4.2957
8	8.2308	-6.1467	-12.6057	-6.5185	23	1.8270	2849.5	-6.6140	9.0966	4.5016	0.26832	-0.09862	-30.8372	-2.5590	-1.9552
9	-4.8995	59.7251	-6.5004	-6.4909	78	-3.9713	3620.8	-6.4514	24.2497	1.5323	2.97495	0.50645	-31.0888	-5.2389	-1.9357
10	-1.1790	11.7382	1.2137	-7.1975	6	-8.1257	2749.4	-6.1885	8.6838	-2.0334	8.54045	-0.18443	-32.6859	-2.5717	-0.1692

Step 2: I regressed the y-variable on the in-sample rows of the x-variables.

Input:



Output:

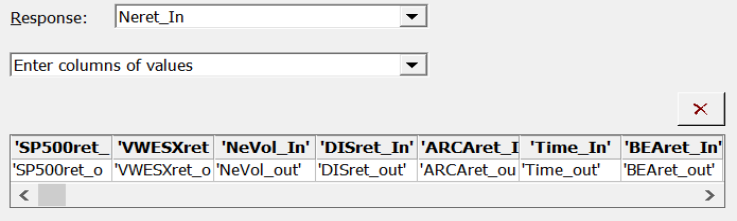
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
10.0104	39.87%	33.92%	25.11%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2.2	24.4	-0.09	0.930	
SP500ret_in	1.749	0.234	7.47	0.000	1.07
VWESXret_In	-0.379	0.341	-1.11	0.270	1.05
NeVol_In	-0.0349	0.0433	-0.81	0.422	1.07
DISret_In	0.200	0.247	0.81	0.420	2.69
ARCAret_In	-0.081	0.562	-0.14	0.886	8.54
Time_In	-0.084	0.159	-0.53	0.597	30.27
BEAret_In	0.215	0.169	1.27	0.206	1.09
CB_In	0.00172	0.00981	0.18	0.861	30.23
NASDAQret_In	-0.255	0.665	-0.38	0.703	10.94

Step 2: To find the y-hat value using Minitab, I used the Stat Predict function and then placed the "out" values under the "in" values:



My resulting "out of sample" y-hat values are shown as follows:

C18	C19	C20	C21	C22
Time_out	BEAret_out	CB_out	NASDAQret_out	PFITS
10	-1.7782	2764.7	-6.4047	5.8891
40	1.5539	3063.2	-4.8153	2.6839
12	-10.3242	2787.9	8.5276	-1.4411
120	-16.1460	4635.6	-1.9741	-6.0670
81	5.0729	3671.8	-2.5691	4.2598
27	0.8084	2931.2	-2.9661	17.1750
56	-0.7840	3253.9	7.7465	-14.5159
52	-0.3034	3216.4	-1.5811	3.5377
94	-4.3365	3957.9	-9.8620	2.1463
48	-13.0093	3264.2	-3.9024	9.0822
108	-14.8228	4308.3	10.9806	-10.9125
18	-5.4151	2823.9	-2.7574	-13.9660
36	-14.0646	3155.9	1.7717	3.1256
80	-5.2831	3646.0	9.0450	-8.1668
72	-10.3266	3486.1	-1.7485	-0.8466

Step 3: I checked the y-hat values given by mini-tab in excel by manually inputting the y-hat equation

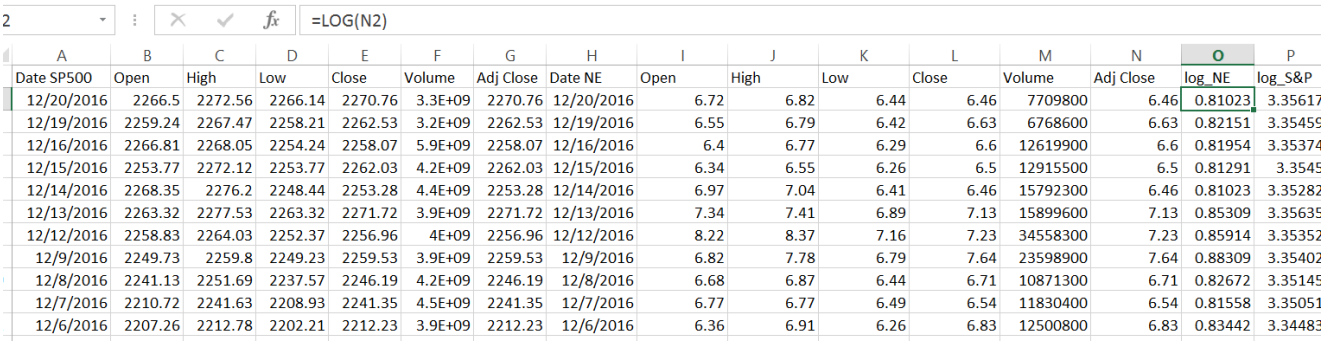
Step 4: Next, I computed the sample standard deviation of the residuals from comparing the out-of-sample y-values to the out-of-sample yhats computed using the in-sample sample slope coefficients

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Neret_out	SP500ret_out	VWESXret_out	NeVol_out	DISret_out	ARCAret_out	Time_out	BEAret_out	CB_out	NASDAQret_out		y-hat	PFITS	Sq_error
2	22.9063051	-8.565734293	-5.569434075	-26.58741	-3.0162914	2.555716669	25	-0.0191504	2886.8	1.173553495		-5.8111	-5.81153	219.656
3	-7.8932731	-0.418858788	1.581593982	-27.55887	9.71952819	2.389050803	96	-10.269478	4015.8	6.814455372		-2.64506	-2.61369	135.833
4	7.02847157	4.32906906	1.01887741	-27.92261	1.72302173	-2.241045224	4	8.47611299	2731	-1.859559206		9.24273	9.24362	0.05431
5	24.0819248	0.852081973	-2.319921828	-28.02221	5.40892639	14.36550689	100	2.84100311	4120.2	13.17667134		2.07625	2.10385	48.0725
6	7.95109012	1.48223383	1.720661321	-29.45152	-3.8163385	-8.532472889	10	-1.7782325	2764.7	-6.404712604		3.99411	3.99885	25.1559
7	-5.571485	1.475922988	3.287615271	-30.44355	-5.6752941	-1.656768551	40	1.55394026	3063.2	-4.815347396		0.82651	0.83623	66.9643
8	8.4020725	-0.862850903	-0.497687793	-30.71016	9.66394915	4.295690449	12	-10.324198	2787.9	8.527616129		3.92572	3.93161	25.8466
9	4.50160772	0.268322495	-0.098619329	-30.83719	-2.5589952	-1.955217419	120	-16.145981	4635.6	-1.974077768		-7.57701	-7.53236	275.118
10	1.53226206	2.974952318	0.506452432	-31.08881	-5.2388701	-1.935710919	81	5.07291839	3671.8	-2.569081685		3.55208	3.57316	29.7854
11	-2.0333704	8.540446162	-0.184430569	-32.68593	-2.5716806	-0.169187197	27	0.80836317	2931.2	-2.966050568		12.4457	12.4546	11.8059
12	-8.0000724	-5.679110746	1.269750447	-32.7111	-5.6880251	8.209671332	56	-0.7840051	3253.9	7.746504212		-3.78155	-3.77022	163.616
13	-5.7211791	2.849538044	2.747168069	-36.2668	-5.3749064	-1.11376563	52	-0.3034212	3216.4	-1.581077755		4.34798	4.36258	21.7314
14	-4.1023268	2.320146079	2.096366876	-36.42975	-7.6541787	-10.61643591	94	-4.3365366	3957.9	-9.862029933		-1.67071	-1.64026	114.071
15	5.45400217	6.530004049	-0.796100295	-37.03834	-7.590942	0.111697117	48	-13.009333	3264.2	-3.902444443		6.62856	6.64771	5.66974
16	-20.497371	-1.753018518	-0.725415173	-39.3364	8.17694032	15.67724258	108	-14.822793	4308.3	10.98055578		-7.76497	-7.7291	281.389
17	2.88247452	-8.59623849	0.605066446	-41.35489	-5.4333374	-2.923949197	18	-5.4151319	2823.9	-2.757448383		-8.53879	-8.53587	307.949
18	-1.4766528	1.777057119	-2.435956249	-42.18388	5.21966843	2.066948954	36	-14.064614	3155.9	1.771737065		3.1028	3.11765	34.8913
19	-1.9762956	-3.129801903	-1.506338223	-43.69165	10.2334001	9.166044456	80	-5.2831069	3646	9.045027934		-3.07973	-3.05927	146.154
20	0.95679153	0.706823046	-0.665342167	-59.94657	-3.4988525	-3.699041586	72	-10.32659	3486.1	-1.748465217		2.27287	2.29561	45.3847
21														
22	coef	17.1	1.312	-0.45	-0.0459	-0.153	-0.47	0.013	0.146	-0.0055	0.763			
23													SSE	1959.15
24													MSE	21.5291
25													S-out	4.63994

Extra Credit #2

"Log Lags"

Step 1: I first computed the log of the daily Range for my stock, NE, and the SP500 for the past ten years in excel



=LOG(N2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Date SP500	Open	High	Low	Close	Volume	Adj Close	Date NE	Open	High	Low	Close	Volume	Adj Close	log_NE	log_S&P
	12/20/2016	2266.5	2272.56	2266.14	2270.76	3.3E+09	2270.76	12/20/2016	6.72	6.82	6.44	6.46	7709800	6.46	0.81023	3.35617
	12/19/2016	2259.24	2267.47	2258.21	2262.53	3.2E+09	2262.53	12/19/2016	6.55	6.79	6.42	6.63	6768600	6.63	0.82151	3.35459
	12/16/2016	2266.81	2268.05	2254.24	2258.07	5.9E+09	2258.07	12/16/2016	6.4	6.77	6.29	6.6	12619900	6.6	0.81954	3.35374
	12/15/2016	2253.77	2272.12	2253.77	2262.03	4.2E+09	2262.03	12/15/2016	6.34	6.55	6.26	6.5	12915500	6.5	0.81291	3.3545
	12/14/2016	2268.35	2276.2	2248.44	2253.28	4.4E+09	2253.28	12/14/2016	6.97	7.04	6.41	6.46	15792300	6.46	0.81023	3.35282
	12/13/2016	2263.32	2277.53	2263.32	2271.72	3.9E+09	2271.72	12/13/2016	7.34	7.41	6.89	7.13	15899600	7.13	0.85309	3.35635
	12/12/2016	2258.83	2264.03	2252.37	2256.96	4E+09	2256.96	12/12/2016	8.22	8.37	7.16	7.23	34558300	7.23	0.85914	3.35352
	12/9/2016	2249.73	2259.8	2249.23	2259.53	3.9E+09	2259.53	12/9/2016	6.82	7.78	6.79	7.64	23598900	7.64	0.88309	3.35402
	12/8/2016	2241.13	2251.69	2237.57	2246.19	4.2E+09	2246.19	12/8/2016	6.68	6.87	6.44	6.71	10871300	6.71	0.82672	3.35145
	12/7/2016	2210.72	2241.63	2208.93	2241.35	4.5E+09	2241.35	12/7/2016	6.77	6.77	6.49	6.54	11830400	6.54	0.81558	3.35051
	12/6/2016	2207.26	2212.78	2202.21	2212.23	3.9E+09	2212.23	12/6/2016	6.36	6.91	6.26	6.83	12500800	6.83	0.83442	3.34483

Step 2: I transferred the logs into Minitab and computed lag columns

	C1	C2	C3	C4	C5	C6
	log_NE	log_S&P	lag_NE_1	lag_NE_2	lag_S&P	RESI
1	0.81023	3.35617	*	*	*	*
2	0.82151	3.35459	0.81023	*	3.35617	*
3	0.81954	3.35374	0.82151	0.81023	3.35459	-0.003619
4	0.81291	3.35450	0.81954	0.82151	3.35374	-0.008708
5	0.81023	3.35282	0.81291	0.81954	3.35450	-0.004933
6	0.85309	3.35635	0.81023	0.81291	3.35282	0.040712
7	0.85914	3.35352	0.85309	0.81023	3.35635	0.005542
8	0.88309	3.35402	0.85914	0.85309	3.35352	0.022310
9	0.82672	3.35145	0.88309	0.85914	3.35402	-0.057342
10	0.81558	3.35051	0.82672	0.88309	3.35145	-0.014906
11	0.83442	3.34483	0.81558	0.82672	3.35051	0.016449
12	0.81291	3.34335	0.83442	0.81558	3.34483	-0.022890
13	0.79029	3.34083	0.81291	0.83442	3.34335	-0.025382

Step 3: I next found the regression of log_NE versus lag_NE_1, lag_NE_2, lag_S&P**500**

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	93.5541	31.1847	174632.10	0.000
lag_NE_1	1	0.4174	0.4174	2337.44	0.000
lag_NE_2	1	0.0004	0.0004	2.49	0.115
lag_S&P	1	0.0003	0.0003	1.42	0.233
Error	2511	0.4484	0.0002		
Total	2514	94.0025			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0133631	99.52%	99.52%	99.52%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.01617	0.00939	1.72	0.085	
lag_NE_1	0.9640	0.0199	48.35	0.000	209.85
lag_NE_2	0.0314	0.0199	1.58	0.115	210.27
lag_S&P	-0.00310	0.00260	-1.19	0.233	1.28

Regression Equation

log_NE = 0.01617 + 0.9640 lag_NE_1 + 0.0314 lag_NE_2 - 0.00310 lag_S&P

Step 4: Analysis

The standard error, S, the average distance between the observed values the regression line, is 0.0133631. As this is a relatively low number, the found lag variables are considered to be good predictors of variables. Furthermore, the r-square value, the measure of proximity between the data and the fitted regression line, is 99.52%, an extremely high proportion. As the r^2 is an estimate of the strength of the relationship between your model and the response variable, the high proportion means that the regression model is a good fit for the data and is a good estimator of future variables.