SDA PROJECT REPORT

STOCK PORTIFOLIO DATASET ANALYSIS

Group members:

R. Navya Sai

B. Devi Neeharika

ABSTRACT:

The main goal of this project is to extract the maximum knowledge from stock portifolio dataset. We apply linear regression model for this dataset and check all the assumptions here.

METHODOLOGY*:*

1. Loading the dataset
2. Check for Null Values.
3. Check for normality of data.
4. Influential points detection.
5. Removal of influential points.
6. Check for correlation between dependent variables (Feature selection)
7. Splitting of data into train and test data and apply regression model.
8. Test of hypothesis based on P value (feature selection).
9. Test of assumptions
   a. Linearity
   b. Homoscedasticity

c. Normality of errors

d. Uncorrelated errors.

10. Goodness of test.

DESCRIPTION OF DATA:

Attribute Information:

The inputs are the weights of the stock-picking concepts as follows

X1=the weight of the Large B/P concept

X2=the weight of the Large ROE concept

X3=the weight of the Large S/P concept

X4=the weight of the Large Return Rate in the last quarter concept

X5=the weight of the Large Market Value concept

X6=the weight of the Small systematic Risk concept

The outputs are the investment performance indicators as follows

Y1=Annual Return

Y2=Excess Return

Y3=Systematic Risk

Y4=Total Risk

Y5=Abs. Win Rate

Y6=Rel. Win Rate

ANALYSIS:

Sample data:

	Large B/P	Large ROE	Large S/P	Large Return Rate in the last quarter	Large Market Value	Small systematic Risk	Annual Return	Excess Return	Systematic Risk	Total Risk	Abs. Win Rate	Rel. Win Rate
0	1.0	0.0	0.0	0.0	0.0	0.0	0.531875	0.478116	0.738015	0.800000	0.52	0.411765
1	0.0	1.0	0.0	0.0	0.0	0.0	0.549712	0.487595	0.571579	0.412231	0.52	0.764706
2	0.0	0.0	1.0	0.0	0.0	0.0	0.692625	0.629895	0.703051	0.756879	0.44	0.376471
3	0.0	0.0	0.0	1.0	0.0	0.0	0.324351	0.255634	0.800000	0.756046	0.36	0.270588
4	0.0	0.0	0.0	0.0	1.0	0.0	0.326615	0.306501	0.432452	0.209289	0.72	0.447059

Summary of data:

	Large B/P	Large ROE	Large S/P	Large Return Rate in the last quarter	Large Market Value	Small systematic Risk	Annual Return	Excess Return	Systematic Risk	Total Risk	Abs. Win Rate	Rel. Win Rate
count	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000	63.000000
mean	0.166619	0.166619	0.166619	0.166619	0.166619	0.166619	0.580151	0.576170	0.426494	0.391749	0.566984	0.547899
variance	0.199304	0.199304	0.199304	0.199304	0.199304	0.199304	0.133358	0.137047	0.118178	0.136653	0.112803	0.159468
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.200000	0.200000	0.200000	0.200000	0.200000	0.200000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.525811	0.519093	0.358600	0.297324	0.520000	0.411765
50%	0.167000	0.167000	0.167000	0.167000	0.167000	0.167000	0.598516	0.587148	0.403418	0.368958	0.560000	0.552941

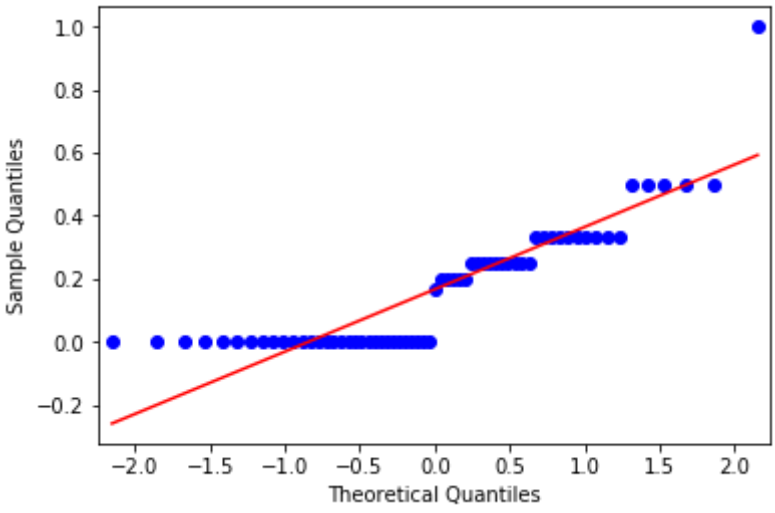
	Large B/P	Large ROE	Large S/P	Large Return Rate in the last quarter	Large Market Value	Small systematic Risk	Annual Return	Excess Return	Systematic Risk	Total Risk	Abs. Win Rate	Rel. Win Rate
75%	0.291500	0.291500	0.291500	0.291500	0.291500	0.291500	0.679636	0.669294	0.470571	0.457749	0.640000	0.694118
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.800000	0.800000	0.800000	0.800000	0.800000	0.800000

Null values:

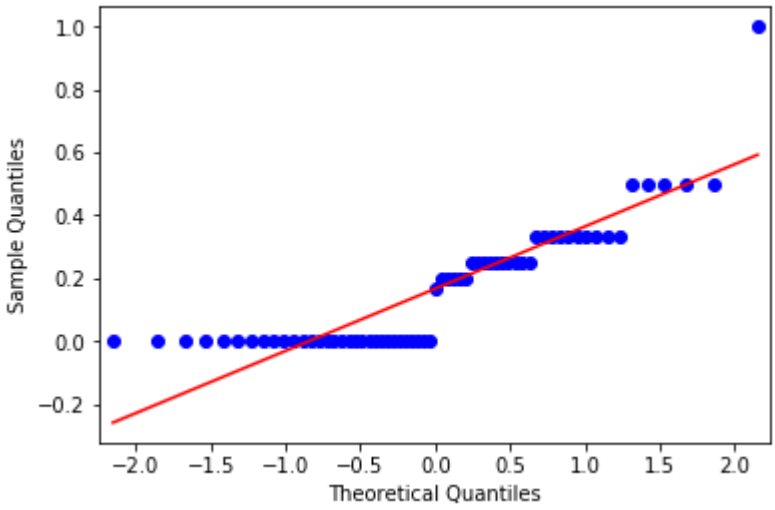
False

Normality of data:

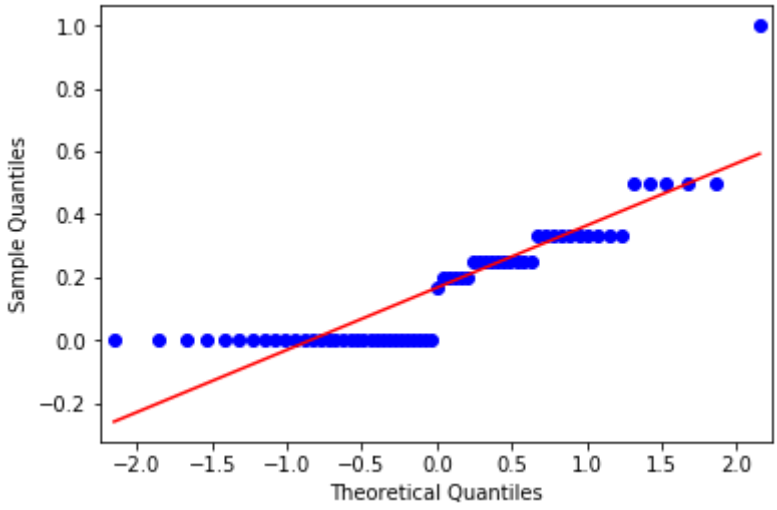
Large B/P



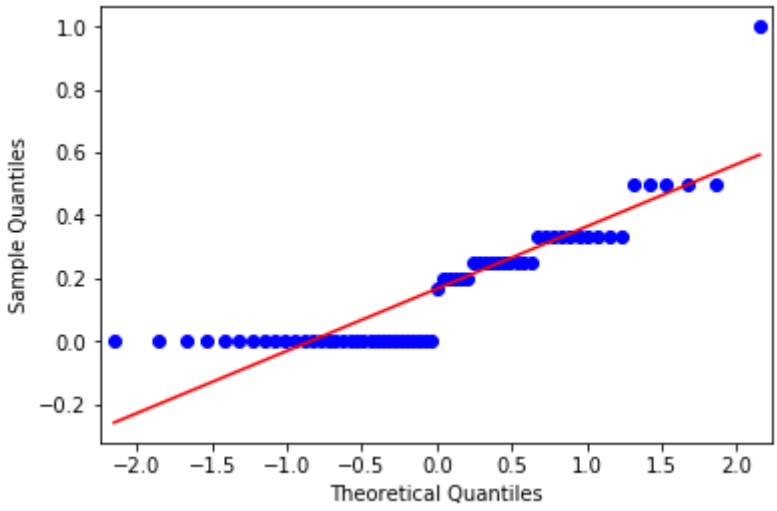
Large ROE



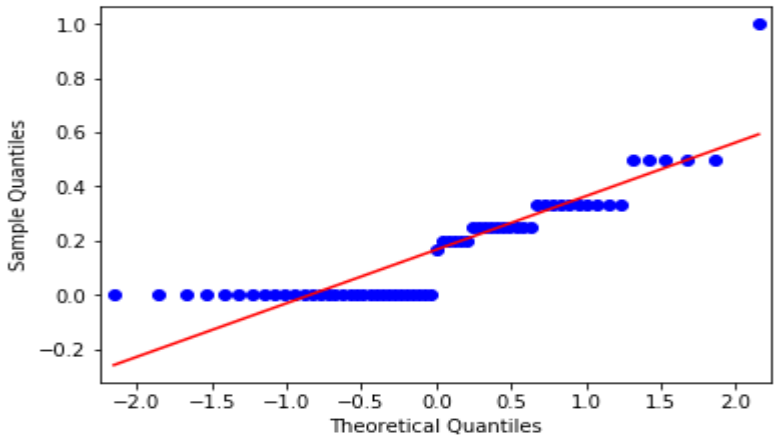
Large S/P



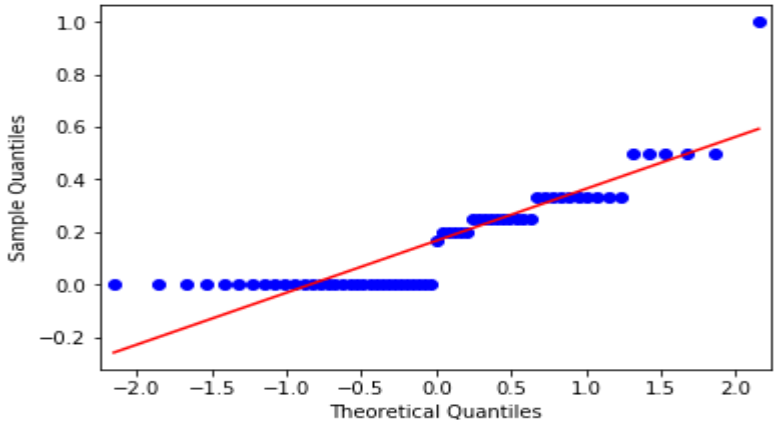
Large Return Rate in the last quarter concept



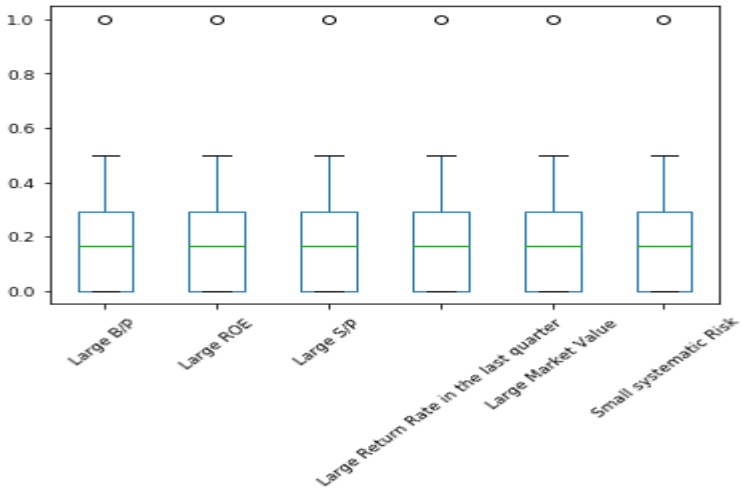
Large Market Value concept



Small systematic Risk concept



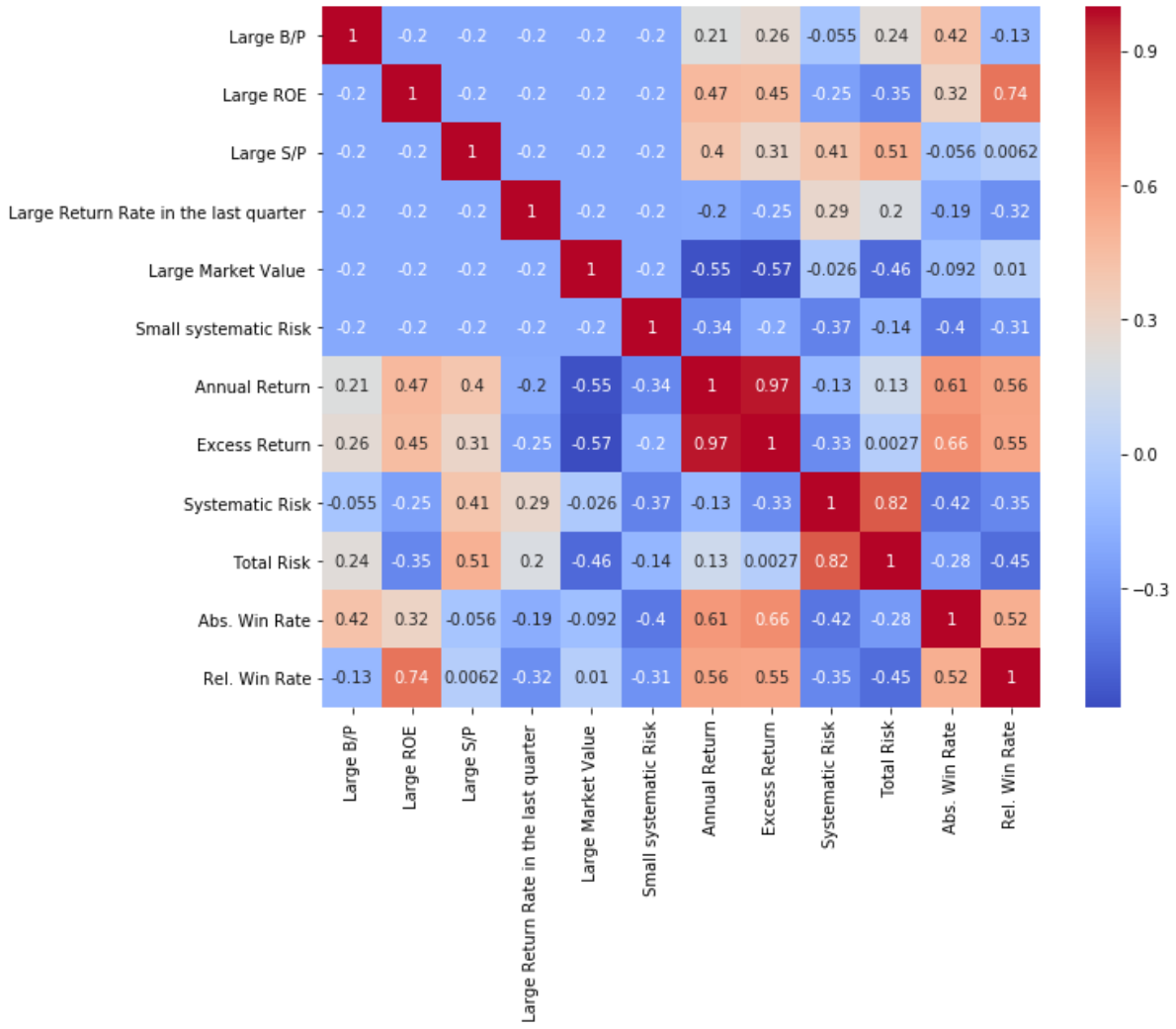
Influential points:



After removing influential points we left with 57 rows.

Data visualization:

Correlation plot:



Correlation plot between the predictors and normalized response variables for the All Period dataset.

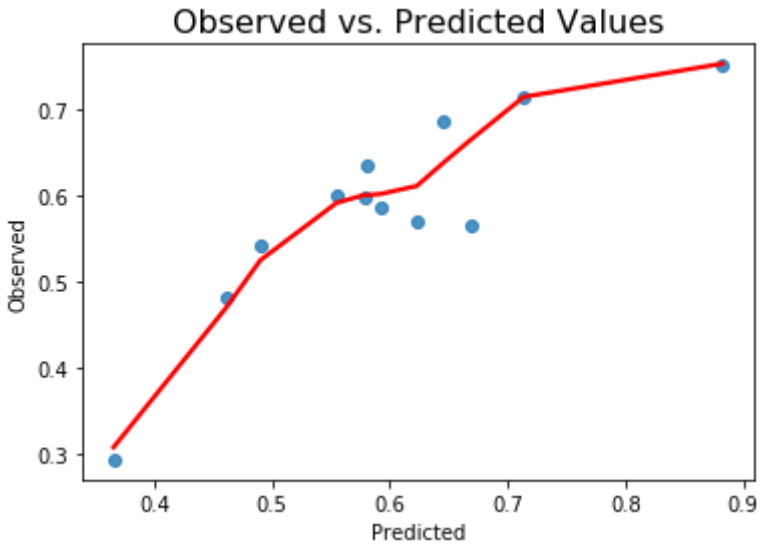
Build a regression model:

Linear regression:

Model for **Annual Return** evaluation parameter

The linear model is: Y = -12.438 + 13.148*large b/p + 13.355*large ROE + 13.286*large s/p+ 12.912*large return rates+ 12.719*large market sales+ 12.817*small system risk

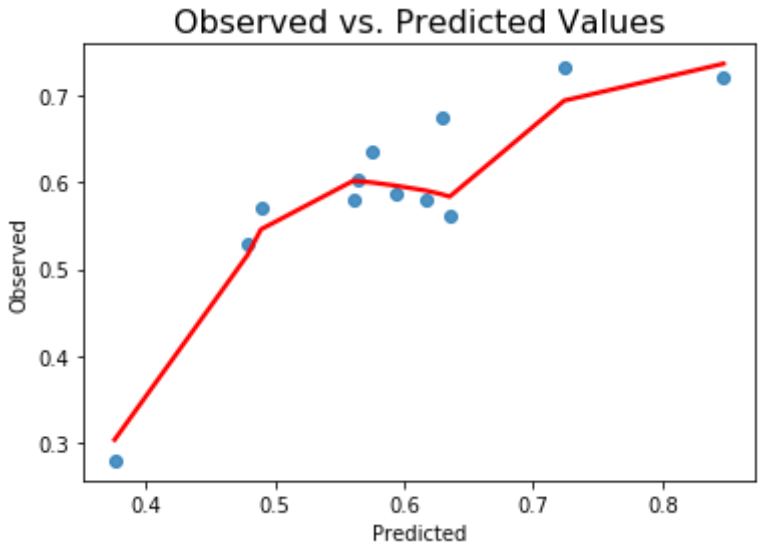
Variance score: 0.7019488999882291



Model for **Excess Return** evaluation parameter

The linear model is: Y = -16.664 + 17.401*large b/p + 17.568*large ROE + 17.455*large s/p+ 17.114*large return rates+ 16.927*large market sales+ 17.128*small system risk

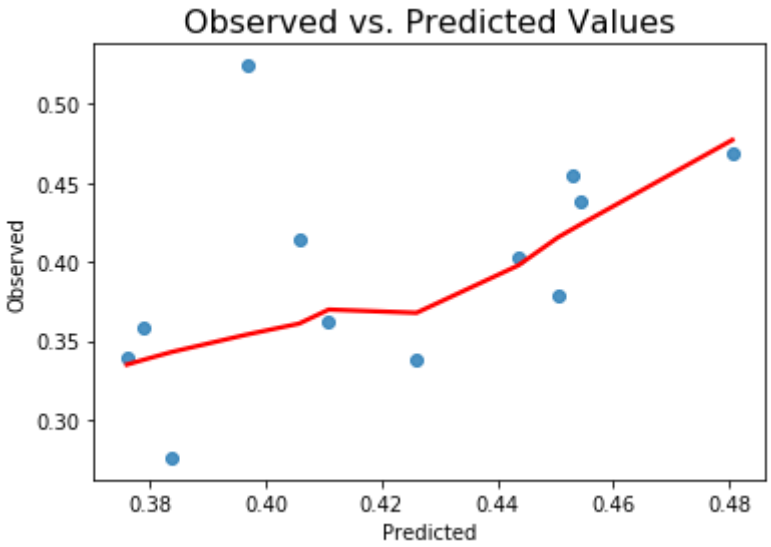
Variance score: 0.6717835969280928



Model for **Systematic Risk** evaluation parameter

The linear model is: Y = 14.279 + -13.893*large b/p + -13.983*large ROE + -13.669*large s/p+ -13.757*large return rates+ -13.875*large market sales+ -14.056*small system risk

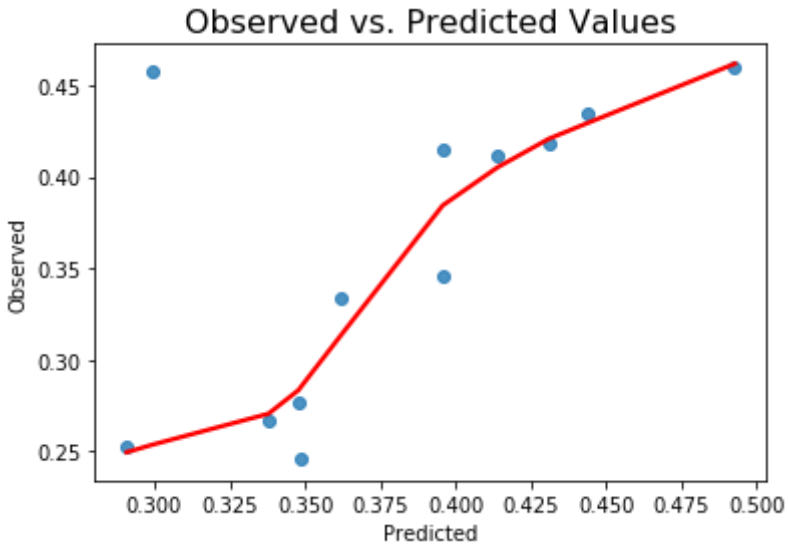
Variance score: 0.08145662584081781



Model for **Total Risk** evaluation parameter

The linear model is: Y = 7.0886 + -6.5722*large b/p + -6.8841*large ROE + -6.431*large s/p+ -6.6196*large return rates+ -6.9572*large market sales+ -6.8111*small system risk

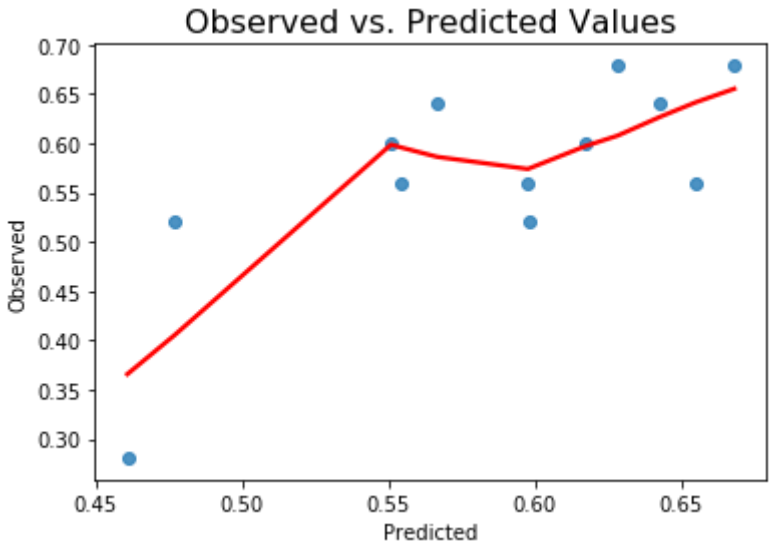
Variance score: 0.30663192303226283



Model for **Abs. Win Rate** evaluation parameter

The linear model is: Y = -20.205 + 20.978*large b/p + 20.954*large ROE + 20.767*large s/p+ 20.718*large return rates+ 20.75*large market sales+ 20.591*small system risk

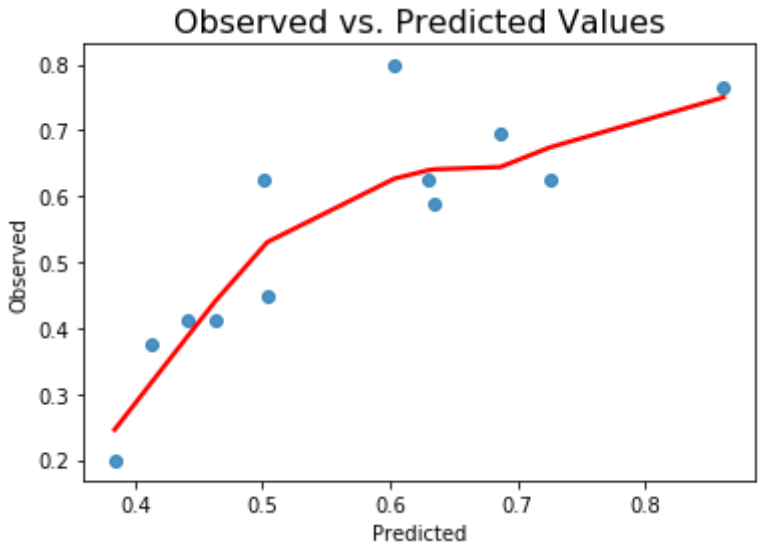
Variance score: 0.5045284905611849



Model for **Rel. Win Rate** evaluation parameter

The linear model is: Y = -46.851 + 47.324*large b/p + 47.982*large ROE + 47.441*large s/p+ 47.204*large return rates+ 47.448*large market sales+ 47.194*small system risk

Variance score: 0.6624426054758186



Ordinary least squares (OLS):

- Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable; the method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted value of the dependent variable configured as a straight line

OLS summary as follows:

```
==============================================================================
================
Dep. Variable:          Annual Return   R-squared (uncentered):
0.990
Model:                            OLS   Adj. R-squared (uncentered):
0.988
Method:                 Least Squares   F-statistic:
644.1
Date:                Thu, 28 Nov 2019   Prob (F-statistic):
2.13e-37
Time:                        10:53:22   Log-Likelihood:
62.057
No. Observations:                  45   AIC:
-112.1
Df Residuals:                      39   BIC:
-101.3
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
==================================
                                                 coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
-----------------------------------
 Large B/P                                     0.7031      0.049     14.35
2      0.000       0.604       0.802
 Large ROE                                     0.9135      0.050     18.35
1      0.000       0.813       1.014
 Large S/P                                     0.8442      0.050     16.97
8      0.000       0.744       0.945
 Large Return Rate in the last quarter         0.4679      0.049      9.49
8      0.000       0.368       0.567
 Large Market Value                            0.2811      0.048      5.89
6      0.000       0.185       0.377
 Small systematic Risk                         0.3766      0.047      7.98
5      0.000       0.281       0.472
==============================================================================
=======
Omnibus:                        3.579   Durbin-Watson:
2.099
```

```
Prob(Omnibus):                  0.167   Jarque-Bera (JB):
3.382
Skew:                          -0.624   Prob(JB):
0.184
Kurtosis:                       2.503   Cond. No.
2.53
================================================================================
=======

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
                            OLS Regression Results
================================================================================
================
Dep. Variable:          Excess Return   R-squared (uncentered):
0.986
Model:                            OLS   Adj. R-squared (uncentered):
0.984
Method:                 Least Squares   F-statistic:
473.7
Date:                Thu, 28 Nov 2019   Prob (F-statistic):
7.88e-35
Time:                        10:53:22   Log-Likelihood:
55.222
No. Observations:                  45   AIC:
-98.44
Df Residuals:                      39   BIC:
-87.60
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
====================================
                                                 coef    std err
t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
------------------------------------
 Large B/P                                     0.7275      0.057     12.75
8      0.000       0.612       0.843
 Large ROE                                     0.8993      0.058     15.52
1      0.000       0.782       1.017
 Large S/P                                     0.7855      0.058     13.57
0      0.000       0.668       0.903
 Large Return Rate in the last quarter         0.4423      0.057      7.71
4      0.000       0.326       0.558
 Large Market Value                            0.2629      0.055      4.73
7      0.000       0.151       0.375
 Small systematic Risk                         0.4608      0.055      8.39
4      0.000       0.350       0.572
================================================================================
=======
Omnibus:                        2.634   Durbin-Watson:
2.016
Prob(Omnibus):                  0.268   Jarque-Bera (JB):
2.378
Skew:                          -0.475   Prob(JB):
0.304
```

```
Kurtosis:                       2.395   Cond. No.
2.53
==============================================================================
=======

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
                                OLS Regression Results
==============================================================================
================
Dep. Variable:        Systematic Risk   R-squared (uncentered):
0.966
Model:                            OLS   Adj. R-squared (uncentered):
0.961
Method:                 Least Squares   F-statistic:
187.0
Date:                Thu, 28 Nov 2019   Prob (F-statistic):
3.78e-27
Time:                        10:53:22   Log-Likelihood:
51.486
No. Observations:                  45   AIC:
-90.97
Df Residuals:                      39   BIC:
-80.13
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
====================================
                                                 coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
-----------------------------------
 Large B/P                                      0.3944      0.062      6.36
5      0.000       0.269       0.520
 Large ROE                                      0.2997      0.063      4.76
1      0.000       0.172       0.427
 Large S/P                                      0.6145      0.063      9.77
1      0.000       0.487       0.742
 Large Return Rate in the last quarter          0.5289      0.062      8.49
0      0.000       0.403       0.655
 Large Market Value                             0.4037      0.060      6.69
5      0.000       0.282       0.526
 Small systematic Risk                          0.2259      0.060      3.78
8      0.001       0.105       0.347
==============================================================================
=======
Omnibus:                        2.416   Durbin-Watson:
2.020
Prob(Omnibus):                  0.299   Jarque-Bera (JB):
1.886
Skew:                           0.502   Prob(JB):
0.389
Kurtosis:                       3.005   Cond. No.
2.53
==============================================================================
=======
```

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

Dep. Variable:	Total Risk	R-squared (uncentered):	0.967
Model:	OLS	Adj. R-squared (uncentered):	0.962
Method:	Least Squares	F-statistic:	192.4
Date:	Thu, 28 Nov 2019	Prob (F-statistic):	2.22e-27
Time:	10:53:22	Log-Likelihood:	54.943
No. Observations:	45	AIC:	-97.89
Df Residuals:	39	BIC:	-87.05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Large B/P	0.5203	0.057	9.068	0.000	0.404	0.636
Large ROE	0.2063	0.058	3.539	0.001	0.088	0.324
Large S/P	0.6599	0.058	11.330	0.000	0.542	0.778
Large Return Rate in the last quarter	0.4725	0.058	8.190	0.000	0.356	0.589
Large Market Value	0.1313	0.056	2.351	0.024	0.018	0.244
Small systematic Risk	0.2791	0.055	5.052	0.000	0.167	0.391

Omnibus:	5.861	Durbin-Watson:	2.038
Prob(Omnibus):	0.053	Jarque-Bera (JB):	4.851
Skew:	0.781	Prob(JB):	0.0884
Kurtosis:	3.381	Cond. No.	2.53

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

```
==============================================================================
================
Dep. Variable:          Abs. Win Rate   R-squared (uncentered):
0.983
Model:                            OLS   Adj. R-squared (uncentered):
0.980
Method:                 Least Squares   F-statistic:
372.4
Date:                Thu, 28 Nov 2019   Prob (F-statistic):
7.94e-33
Time:                        10:53:22   Log-Likelihood:
51.463
No. Observations:                  45   AIC:
-90.93
Df Residuals:                      39   BIC:
-80.09
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
====================================
                                                 coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
-----------------------------------
 Large B/P                                     0.7621      0.062     12.29
4      0.000       0.637       0.887
 Large ROE                                     0.7432      0.063     11.79
8      0.000       0.616       0.871
 Large S/P                                     0.5549      0.063      8.81
8      0.000       0.428       0.682
 Large Return Rate in the last quarter         0.5027      0.062      8.06
5      0.000       0.377       0.629
 Large Market Value                            0.5454      0.060      9.04
2      0.000       0.423       0.667
 Small systematic Risk                         0.3820      0.060      6.40
0      0.000       0.261       0.503
==============================================================================
=======
Omnibus:                        0.405   Durbin-Watson:
1.902
Prob(Omnibus):                  0.817   Jarque-Bera (JB):
0.067
Skew:                           0.082   Prob(JB):
0.967
Kurtosis:                       3.095   Cond. No.
2.53
==============================================================================
=======

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
                            OLS Regression Results
==============================================================================
================
Dep. Variable:          Rel. Win Rate   R-squared (uncentered):
0.979
```

```
Model:                            OLS   Adj. R-squared (uncentered):
0.975
Method:                 Least Squares   F-statistic:
297.5
Date:                Thu, 28 Nov 2019   Prob (F-statistic):
5.76e-31
Time:                        10:53:22   Log-Likelihood:
46.904
No. Observations:                  45   AIC:
-81.81
Df Residuals:                      39   BIC:
-70.97
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
====================================
                                            coef    std err
t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------
 Large B/P                                  0.4473      0.069      6.52
1      0.000       0.309       0.586
 Large ROE                                  1.1190      0.070     16.05
3      0.000       0.978       1.260
 Large S/P                                  0.5754      0.070      8.26
3      0.000       0.435       0.716
 Large Return Rate in the last quarter      0.3307      0.069      4.79
4      0.000       0.191       0.470
 Large Market Value                         0.5973      0.067      8.94
8      0.000       0.462       0.732
 Small systematic Risk                      0.3330      0.066      5.04
2      0.000       0.199       0.467
==============================================================================
=======
Omnibus:                        2.566   Durbin-Watson:
1.945
Prob(Omnibus):                  0.277   Jarque-Bera (JB):
2.134
Skew:                           0.532   Prob(JB):
0.344
Kurtosis:                       2.920   Cond. No.
2.53
==============================================================================
=======

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

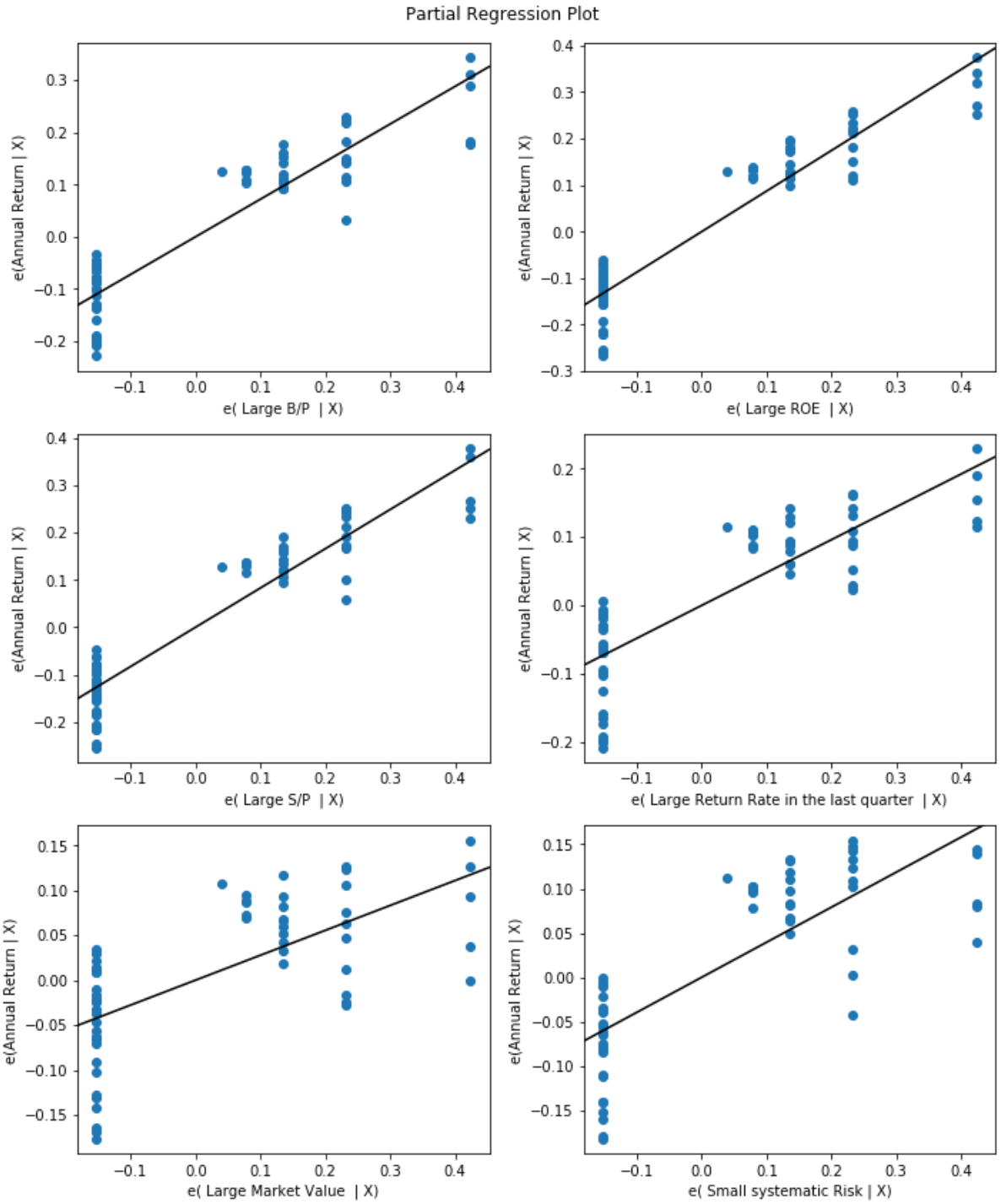
- If **p> 0.05** , we fail to reject null hypothesis otherwise we reject null hypothesis.

- After observing p values if **p value > 0.05** we assume that features are less contributed to evaluation parameters. After observing P values in the above summary all are below 0.05 so we can assume that all are important for predicting evaluation parameters.

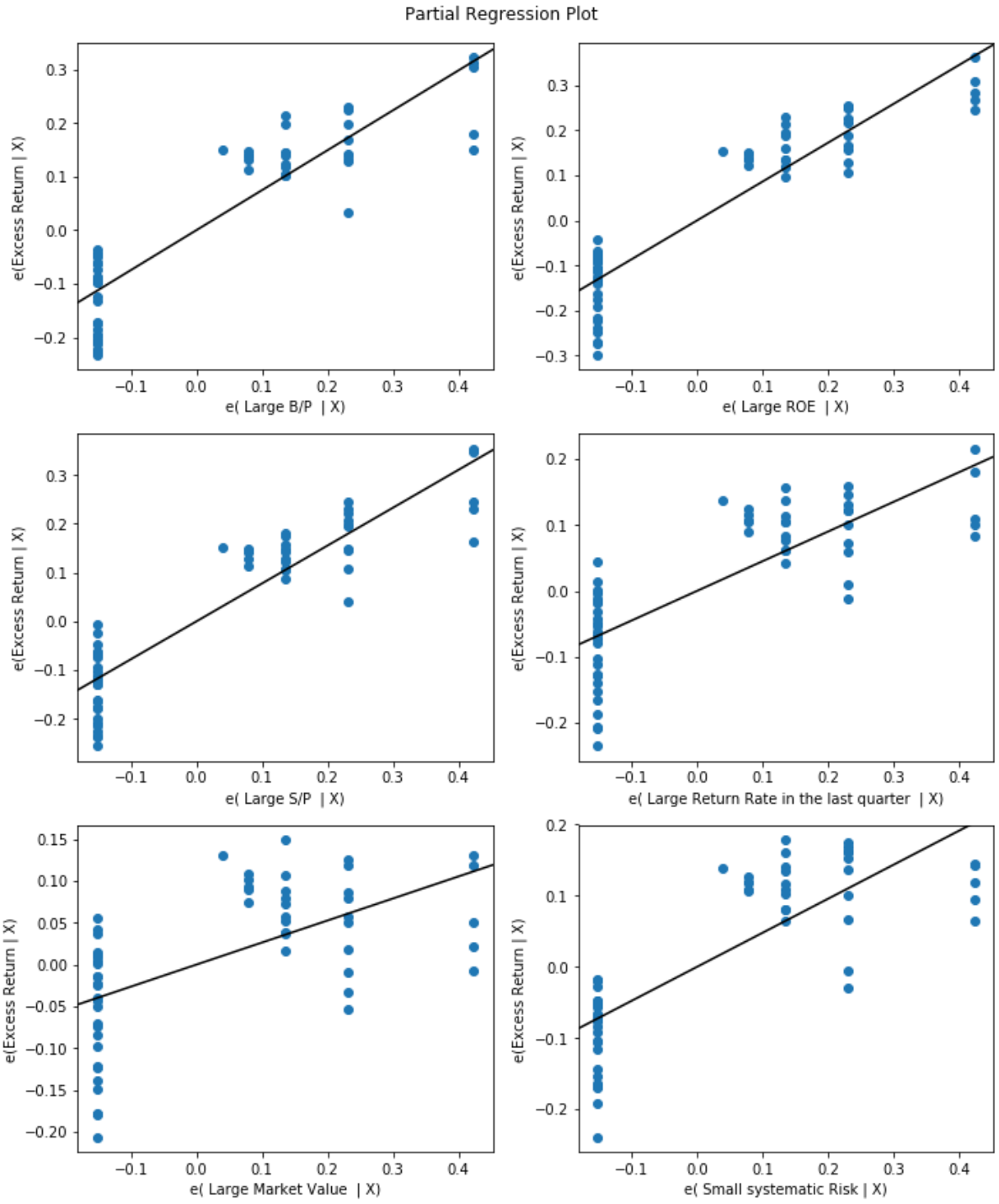
Test of assumptions:

LINEARITY:

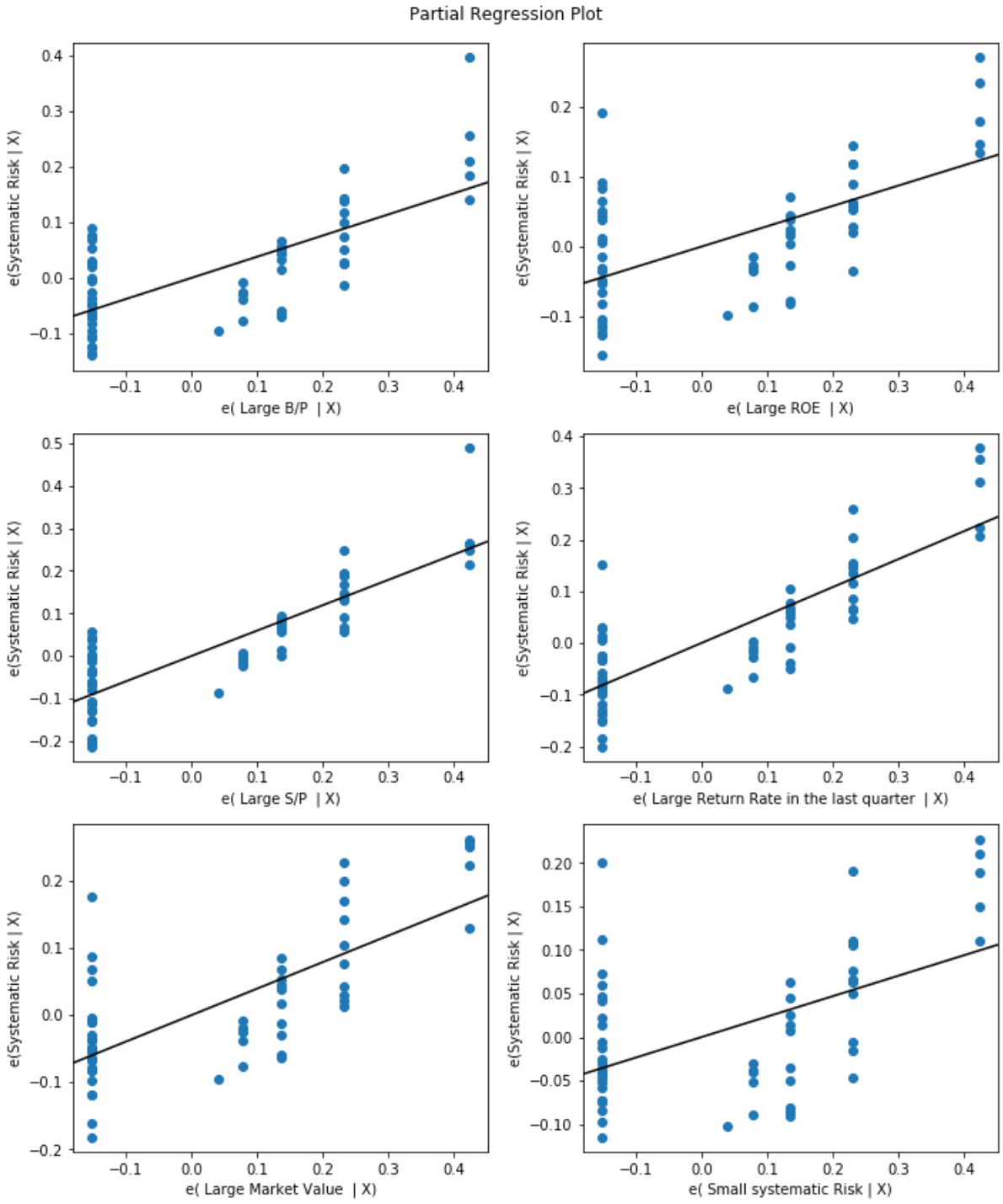
Annual return


Partial Regression Plot
e(Annual Return | X)
e(Large B/P | X)
e(Large ROE | X)
e(Large S/P | X)
e(Large Return Rate in the last quarter | X)
e(Large Market Value | X)
e(Small systematic Risk | X)

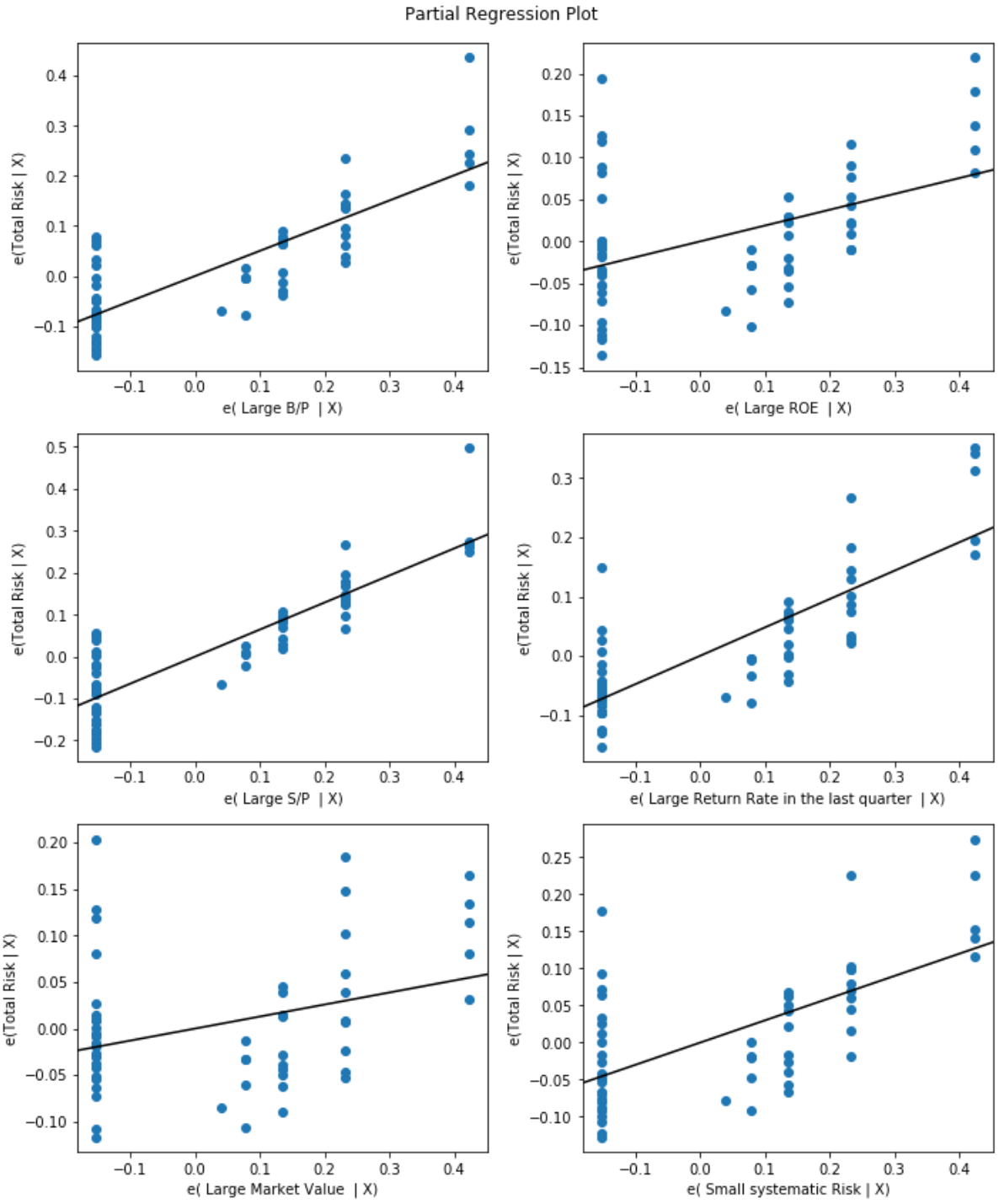
Excess return



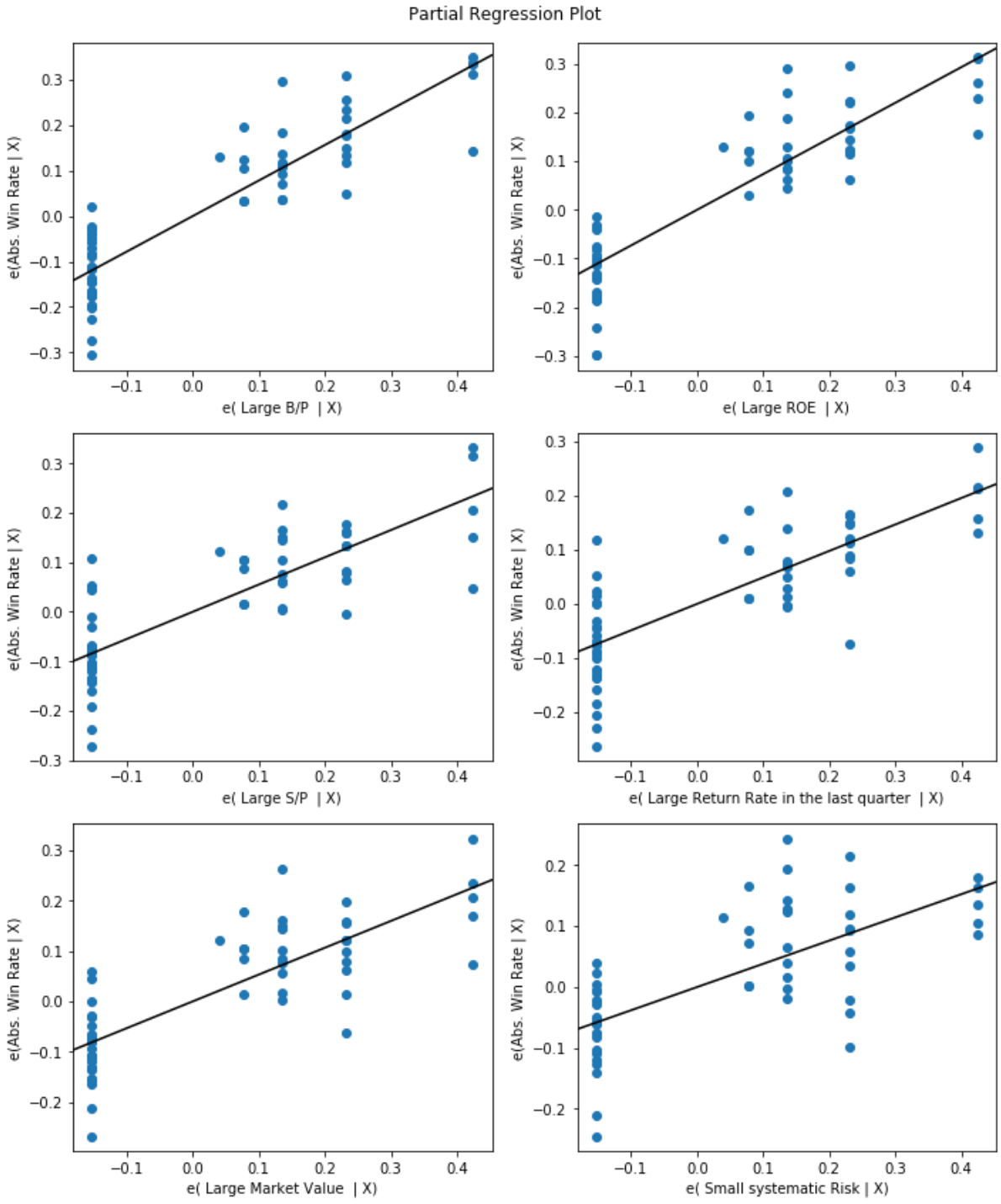
Systematic risk



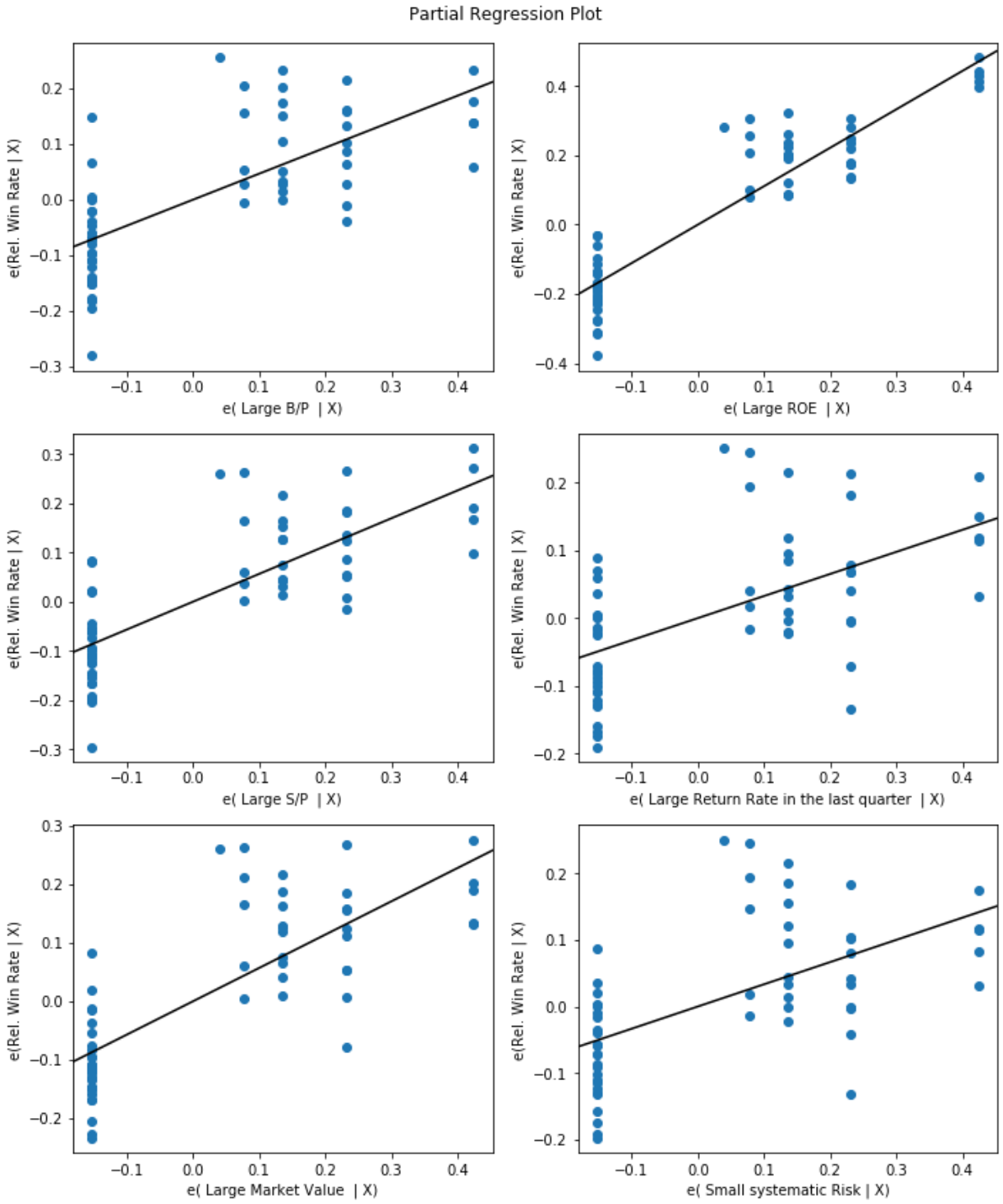
Total risk


Partial Regression Plot
e(Total Risk | X)
e(Large B/P | X)
e(Large ROE | X)
e(Large S/P | X)
e(Large Return Rate in the last quarter | X)
e(Large Market Value | X)
e(Small systematic Risk | X)

Abs. win rate


Partial Regression Plot
e(Abs. Win Rate | X)
e(Large B/P | X)
e(Large ROE | X)
e(Large S/P | X)
e(Large Return Rate in the last quarter | X)
e(Large Market Value | X)
e(Small systematic Risk | X)

Rel. win rate

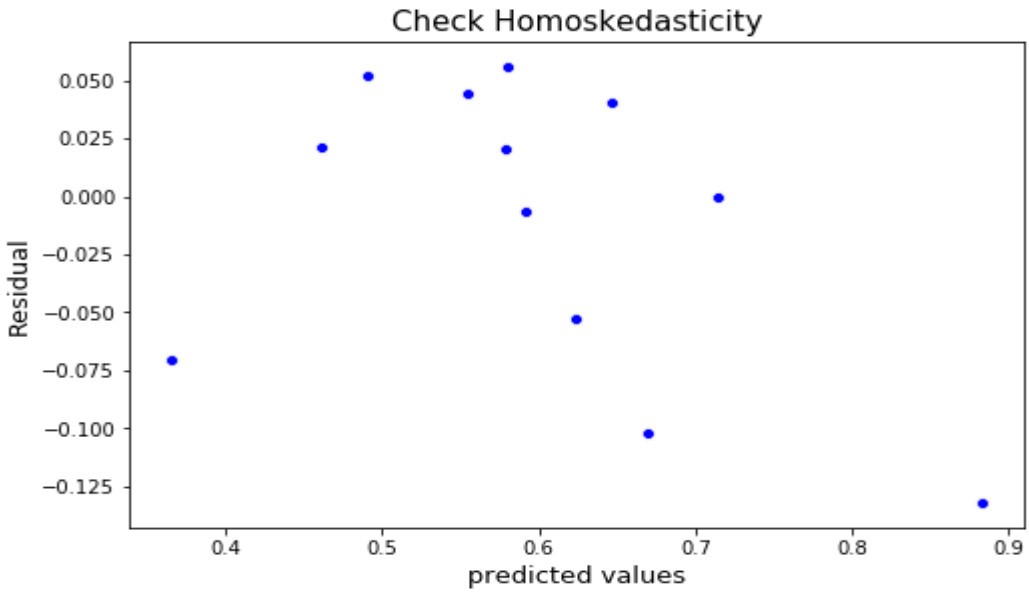


If we observe carefully, all the partial residual plots between the independent variable and dependent variable are linear. Linearity condition is satisfied.

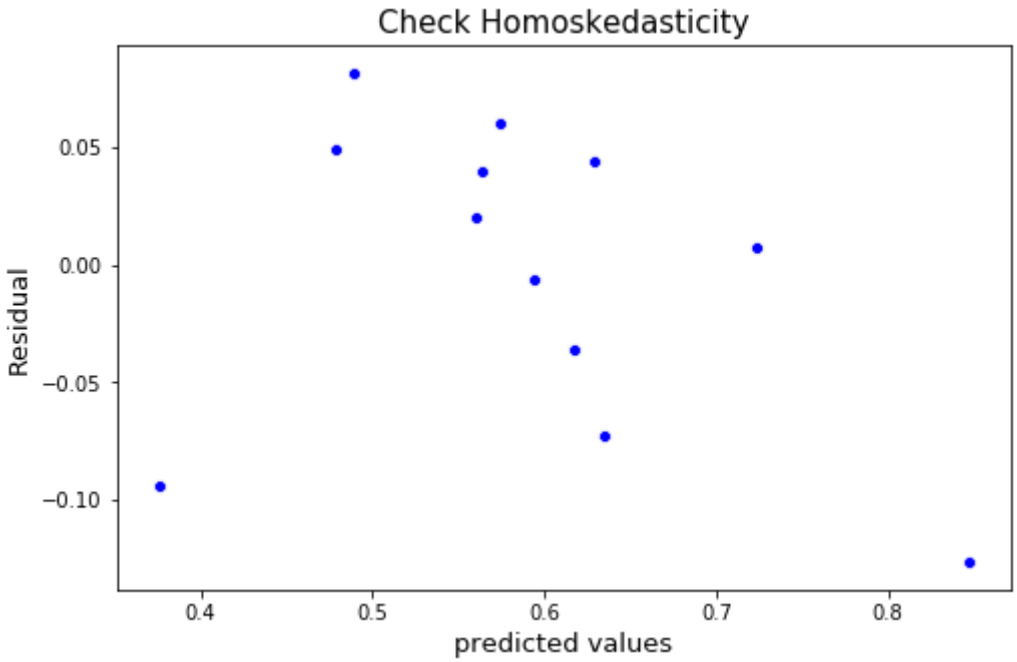
HOMOSCEDASTICITY:

- To check homoscedasticity, we plot the residuals vs predicted values/fitted values.
- If we see any kind of funnel shape, we can say that there is heteroscedasticity.

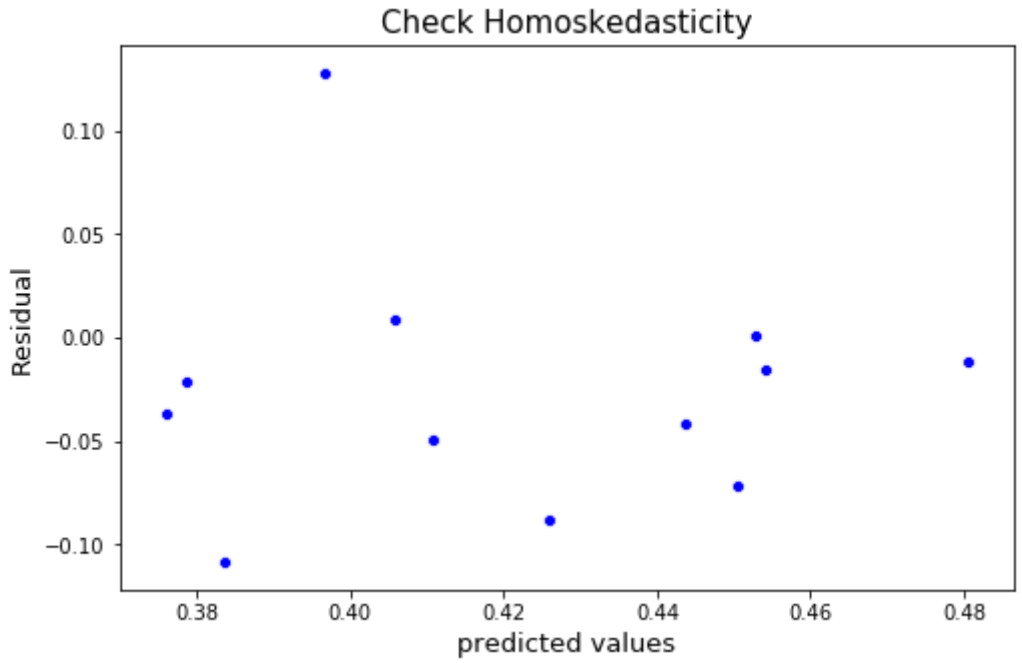
Annual return



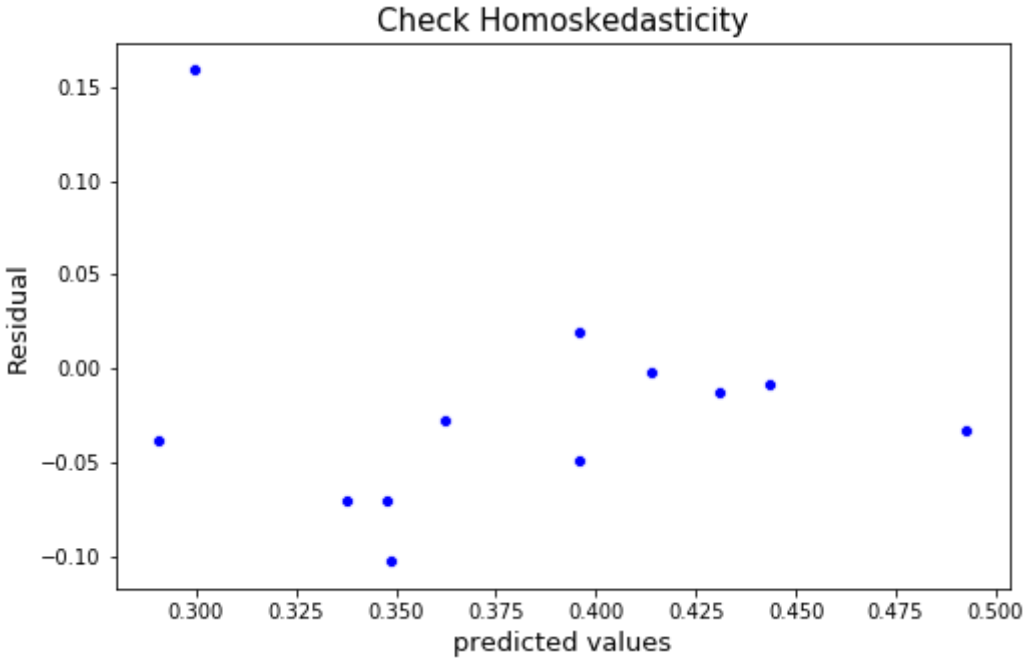
Excess return



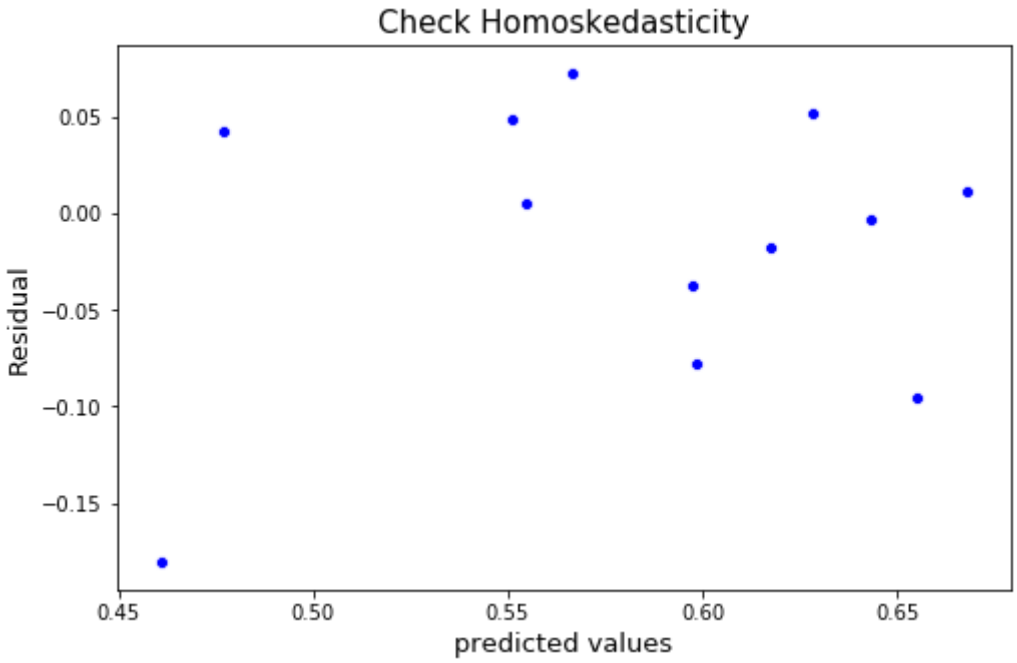
Systematic risk



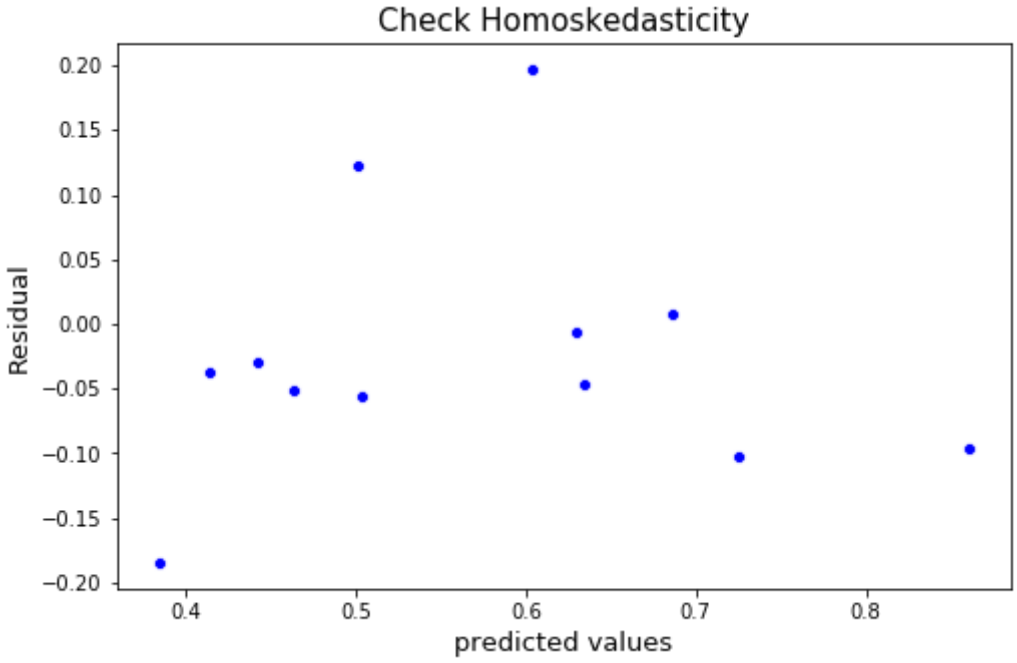
Total risk



Abs. win rate



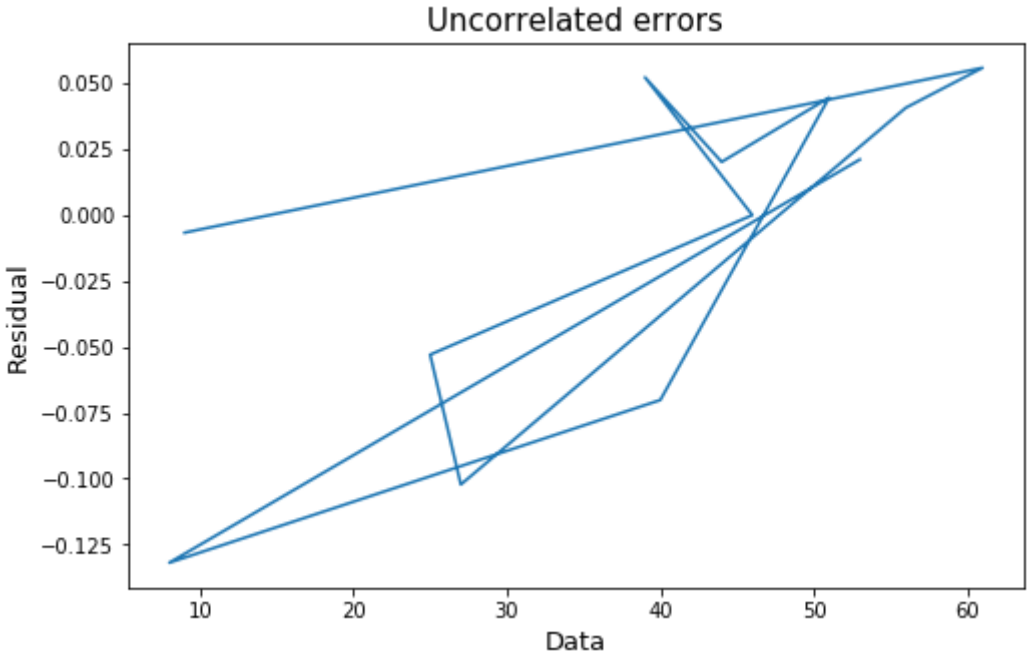
Rel. win rate



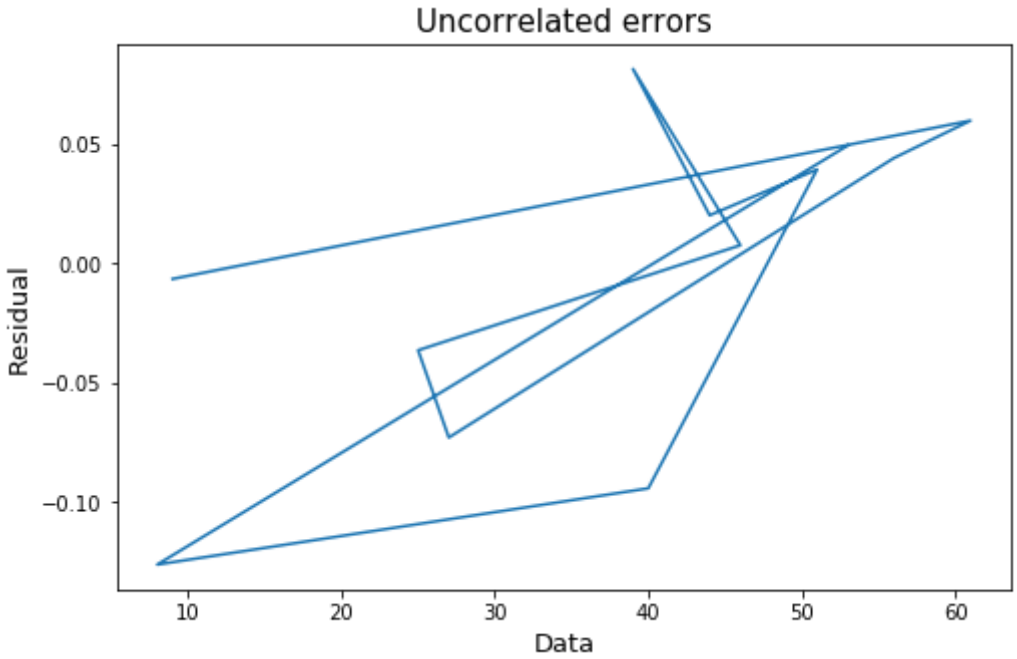
- The points are random. which confirms that there is homoscedasticity.
- It means that the variance of Y across all X is same. - We can conclude that, Homoscedasticity condition hold in this case.

Uncorrelated errors

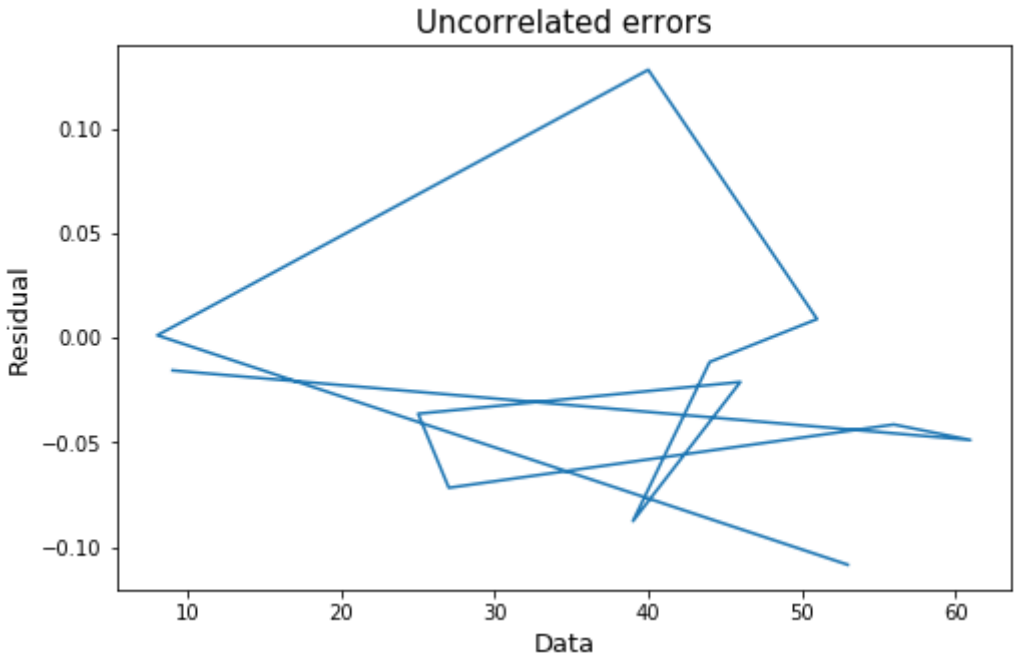
Annual return



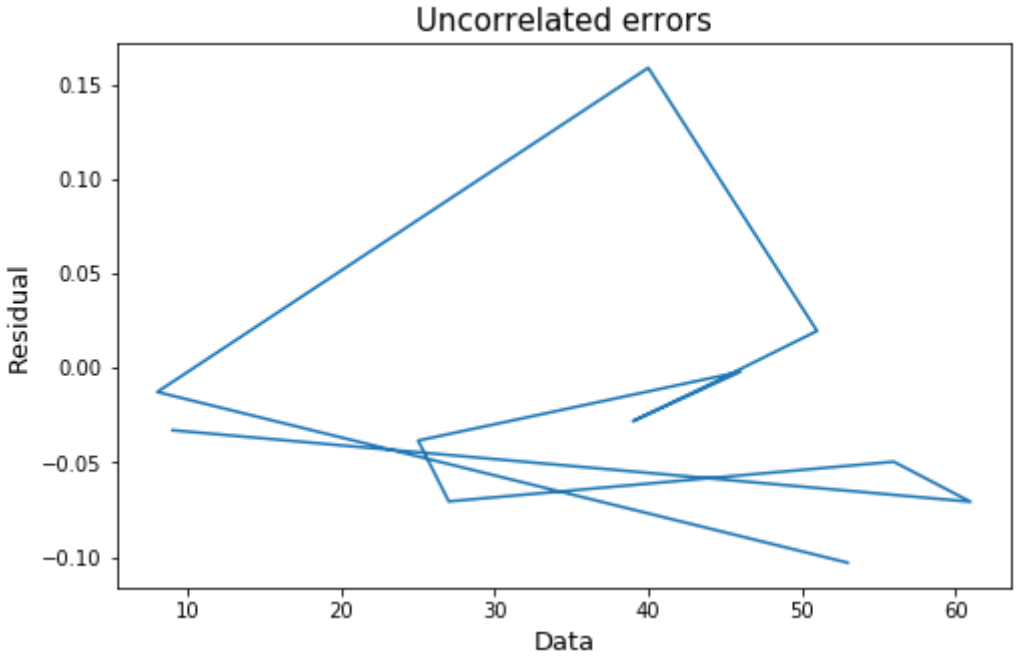
Excess return



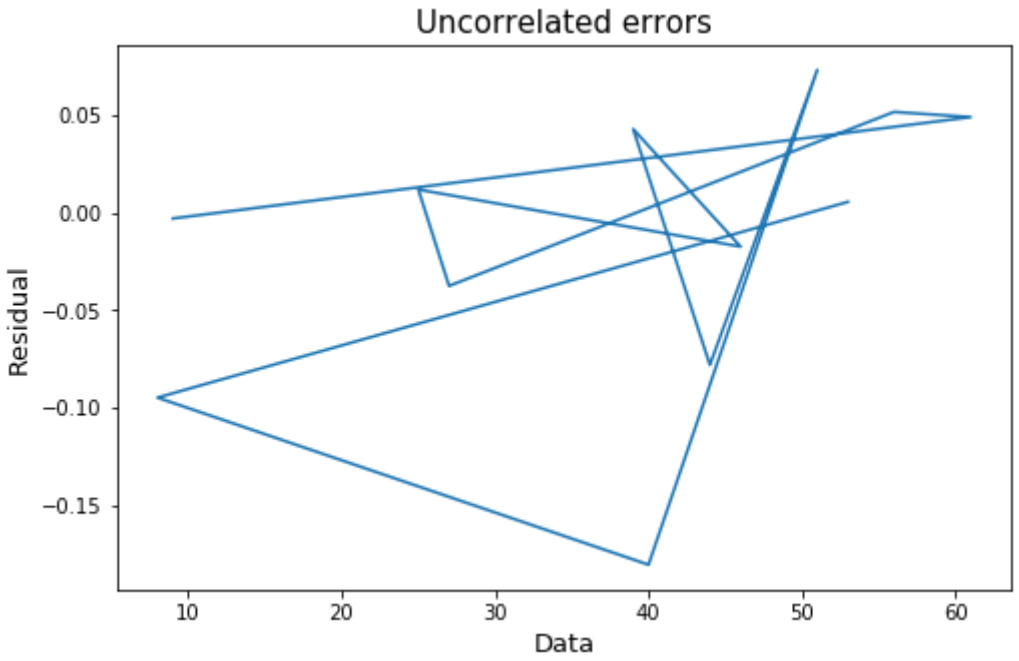
Systematic risk



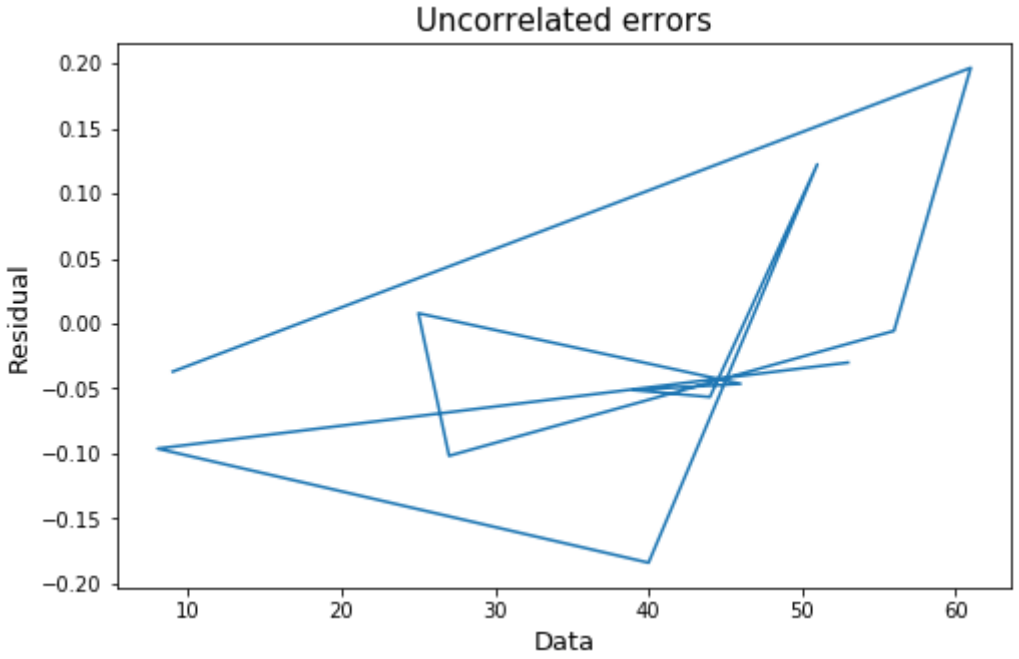
Total risk



Abs. win rate



Rel. win rate



- If we observe, there exists correlation/pattern between errors.
- We can also check this condition using the Durbin-Watson test:
  - If DW = 2, then there is no correlation.
  - If DW < 2, then the errors are positively correlated.
  - If DW > 2, then the errors are negatively correlated.

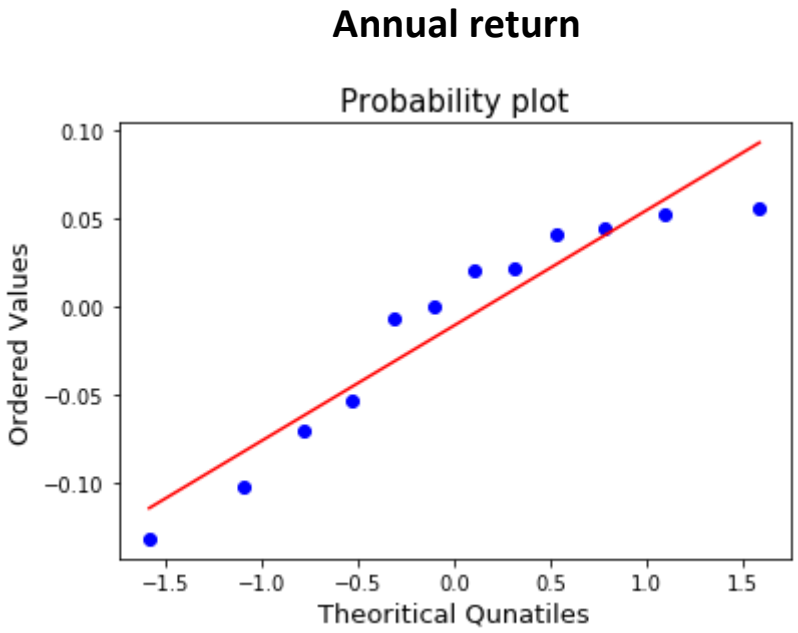
- If we perform Durbin-Watson test, the values of DW are

❖ 1.6215799780769184
❖ 1.6819424442764341
❖ 1.1953060659793757
❖ 1.215215373547569
❖ 2.2147484787943914
❖ 2.1946702508706504

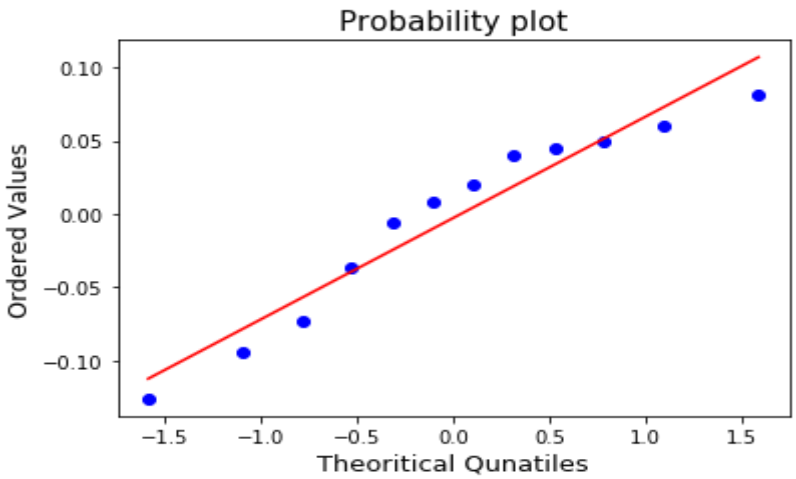
- According to the test, we can say that for first 4 evaluation parameters errors are positively correlated and remaining are negatively correlated.
- However, this is a point estimate for perfect un-correlation of errors (DW=2). So, we won't get DW as 2 on real data. If it around 2, then we can conclude that the errors are uncorrelated.

Normality of error terms:

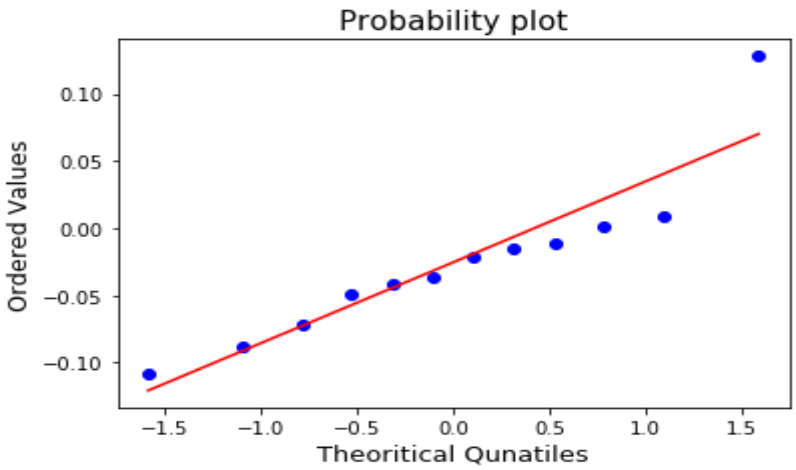
- This can be checked by plotting probability-probability plot (p-p plot) or Quantile-Quantile plot(Q-Q plot).



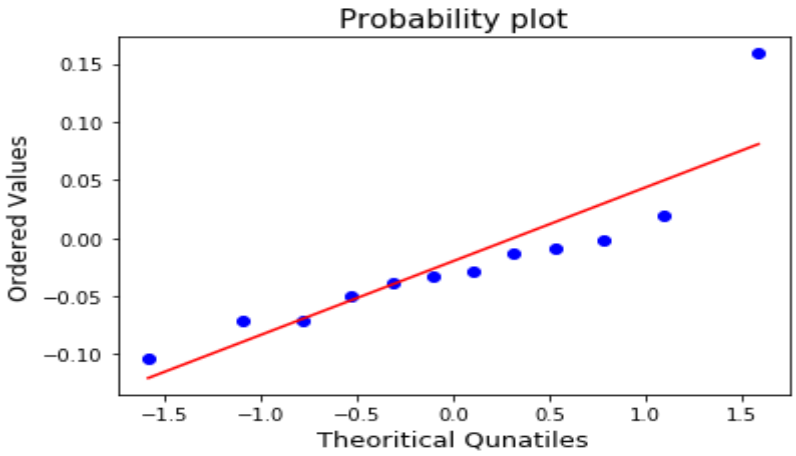
Excess return



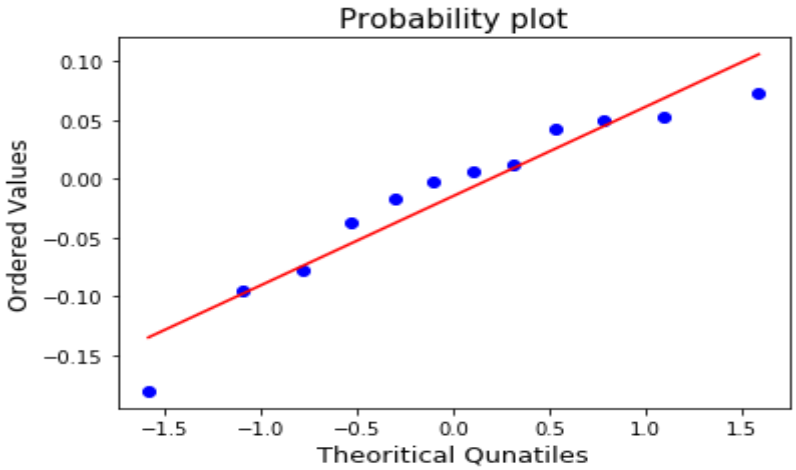
Systematic risk



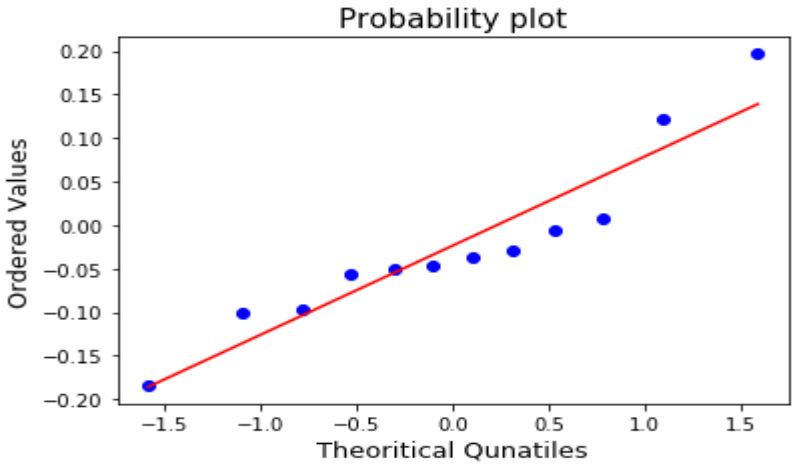
Total risk



Abs. win rate



Rel. win rate



- If we observe the above plots, we can conclude that the errors are following a Normal distribution, because the

plot shows the fluctuation around the line and there is not much deviation. The graphs are linear.

RESULTS:

R-square values of all evaluation parameters w.r.t models

Model	Annual return	Excess return	Systematic risk	Total risk	Absolute win rate	Rel.win rate
Linear regression	0.701	0.671	0.081	0.306	0.504	0.662
svm	0.746	0.688	0.190	0.226	0.296	0.514
Linear regression ridge	0.682	0.625	0.103	0.297	0.354	0.4795
Linear regression lasso	-0.011	-0.007	-0.034	-0.05	-0.010	-0.011
Decision tree	0.05	-0.088	-1.272	0.125	0.476	0.580