22b4217-assignment-5

September 2, 2023

[43]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

[44]:
```
df = pd.read_csv('MLR-Feature-Elimination.csv')
df.head()
```

[44]:
```
      c1  c2         c26         c27        c28         c29        c30  \
0  43344   2  493.796764  104.553871  41.187601  290.965340  14.379552
1  43345   2  493.661889  104.513206  41.580752  290.621190  14.315323
2  43346   2  495.644947  104.502457  40.744572  292.152424  14.566180
3  43347   2  494.354041  104.452871  40.288181  292.676229  14.605181
4  43348   2  492.051373  104.488584  41.266692  289.017462  14.548926

         c31        c32        c33  …        c19       c20       c21  \
0  71.731990  48.679005 -69.203403  …  13.599070  7.120964  9.257515
1  78.599820  48.057417 -69.414081  …  13.167193  7.793413  9.218110
2  78.832458  47.320586 -69.645378  …  12.611031  7.289157  9.599612
3  72.736626  47.980460 -69.452794  …  14.832367  7.958076  9.436385
4  76.621067  48.217299 -69.344057  …  15.943873  8.757605  9.954739

        c22        c23       c34       c35       c36       c52      c241
0  2.743170  44.703468  0.147467  1.454651  0.049850  7.870521  2.184083
1  2.596314  43.973557  0.225583  1.457910  0.049859  7.897945  2.233879
2  2.557701  43.966172  0.197137  1.461920  0.049648  7.609317  2.088296
3  2.897314  43.154569  0.168861  1.490899  0.049995  8.095649  2.089270
4  2.917772  43.044778  0.244714  1.473343  0.049860  7.739171  2.096676

[5 rows x 41 columns]
```

[45]:
```
#Copying the Data to the respective Variables

Y=df['c52']
Temp_X=df
X=Temp_X.drop('c52',axis=1)
del Temp_X
print(X.head())
```

```
        c1  c2         c26         c27        c28         c29        c30  \
0  43344   2  493.796764  104.553871  41.187601  290.965340  14.379552
1  43345   2  493.661889  104.513206  41.580752  290.621190  14.315323
2  43346   2  495.644947  104.502457  40.744572  292.152424  14.566180
3  43347   2  494.354041  104.452871  40.288181  292.676229  14.605181
4  43348   2  492.051373  104.488584  41.266692  289.017462  14.548926

         c31        c32        c33  …        c17        c19       c20  \
0  71.731990  48.679005 -69.203403  …  28.334700  13.599070  7.120964
1  78.599820  48.057417 -69.414081  …  28.211453  13.167193  7.793413
2  78.832458  47.320586 -69.645378  …  28.949064  12.611031  7.289157
3  72.736626  47.980460 -69.452794  …  33.964274  14.832367  7.958076
4  76.621067  48.217299 -69.344057  …  36.744817  15.943873  8.757605

        c21       c22        c23       c34       c35       c36      c241
0  9.257515  2.743170  44.703468  0.147467  1.454651  0.049850  2.184083
1  9.218110  2.596314  43.973557  0.225583  1.457910  0.049859  2.233879
2  9.599612  2.557701  43.966172  0.197137  1.461920  0.049648  2.088296
3  9.436385  2.897314  43.154569  0.168861  1.490899  0.049995  2.089270
4  9.954739  2.917772  43.044778  0.244714  1.473343  0.049860  2.096676

[5 rows x 40 columns]
```

[46]:
```python
import statsmodels.api as sm
```

[47]:
```python
#Here,we don't need to add any constants as we are already having a column of constants (c2)
mlr_model = sm.OLS(Y, X).fit()
print(mlr_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    c52   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.787
Method:                 Least Squares   F-statistic:                     97.90
Date:                Sat, 02 Sep 2023   Prob (F-statistic):          6.30e-308
Time:                        22:10:28   Log-Likelihood:                -1454.7
No. Observations:                1025   AIC:                             2989.
Df Residuals:                     985   BIC:                             3187.
Df Model:                          39
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c1             0.0006      0.000      1.191      0.234      -0.000       0.001
c2           -98.3305     52.926     -1.858      0.063    -202.191       5.530
c26            0.3737      0.048      7.828      0.000       0.280       0.467
c27           -0.1454      0.876     -0.166      0.868      -1.865       1.574
```

c28	0.1911	0.045	4.270	0.000	0.103	0.279
c29	-0.4390	0.048	-9.165	0.000	-0.533	-0.345
c30	3.5436	0.466	7.606	0.000	2.629	4.458
c31	0.2643	0.034	7.800	0.000	0.198	0.331
c32	0.0871	0.195	0.447	0.655	-0.295	0.470
c33	-0.4894	0.454	-1.079	0.281	-1.379	0.401
c39	16.9297	1.573	10.763	0.000	13.843	20.017
c139	-0.8595	0.221	-3.896	0.000	-1.292	-0.427
c142	-0.2540	0.078	-3.260	0.001	-0.407	-0.101
c143	-0.2058	0.039	-5.285	0.000	-0.282	-0.129
c155	-0.0478	0.014	-3.461	0.001	-0.075	-0.021
c157	0.2445	0.042	5.803	0.000	0.162	0.327
c158	0.3032	0.026	11.616	0.000	0.252	0.354
c160	0.0043	0.002	2.415	0.016	0.001	0.008
c161	0.0105	0.001	9.886	0.000	0.008	0.013
c162	0.0022	0.002	1.318	0.188	-0.001	0.005
c163	0.0062	0.002	2.887	0.004	0.002	0.010
c7	0.2134	0.287	0.744	0.457	-0.349	0.776
c8	-0.5668	0.136	-4.176	0.000	-0.833	-0.300
c9	-0.7695	0.075	-10.303	0.000	-0.916	-0.623
c10	10.6356	1.560	6.817	0.000	7.574	13.697
c11	0.2940	0.078	3.747	0.000	0.140	0.448
c12	-0.1646	0.109	-1.509	0.132	-0.379	0.049
c13	0.0439	0.051	0.861	0.390	-0.056	0.144
c15	-0.3619	0.061	-5.905	0.000	-0.482	-0.242
c16	-0.4196	0.102	-4.128	0.000	-0.619	-0.220
c17	-0.0929	0.021	-4.427	0.000	-0.134	-0.052
c19	0.3981	0.213	1.872	0.062	-0.019	0.815
c20	0.1970	0.041	4.791	0.000	0.116	0.278
c21	-0.2064	0.049	-4.213	0.000	-0.302	-0.110
c22	-0.0848	0.036	-2.349	0.019	-0.156	-0.014
c23	-0.2973	0.047	-6.274	0.000	-0.390	-0.204
c34	-0.4606	1.799	-0.256	0.798	-3.991	3.070
c35	7.1614	1.592	4.499	0.000	4.038	10.285
c36	3.0892	88.338	0.035	0.972	-170.263	176.441
c241	13.0944	1.898	6.899	0.000	9.370	16.819

Omnibus:	32.850	Durbin-Watson:	0.571
Prob(Omnibus):	0.000	Jarque-Bera (JB):	81.143
Skew:	-0.027	Prob(JB):	2.40e-18
Kurtosis:	4.377	Cond. No.	1.22e+08

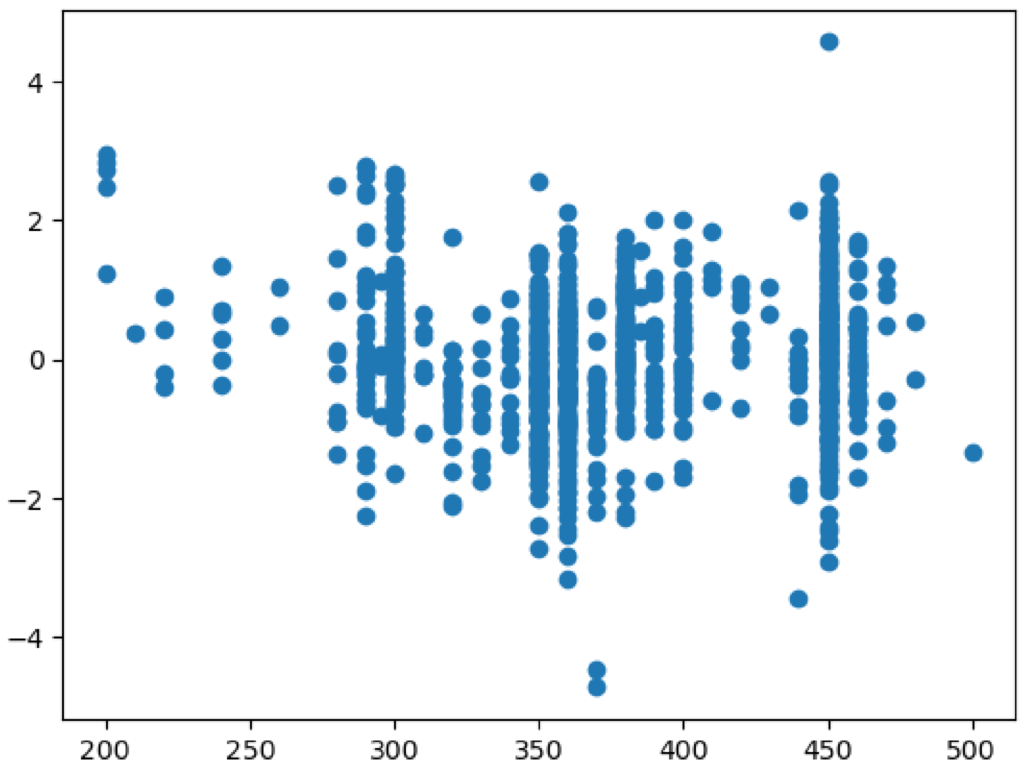
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.22e+08. This might indicate that there are strong multicollinearity or other numerical problems.

[49]:
```
#Prediction of Y
y_predicted=mlr_model.predict(X)
y_predicted.head()
```

[49]:
```
0    7.710503
1    8.167507
2    7.867975
3    6.344570
4    5.990533
dtype: float64
```

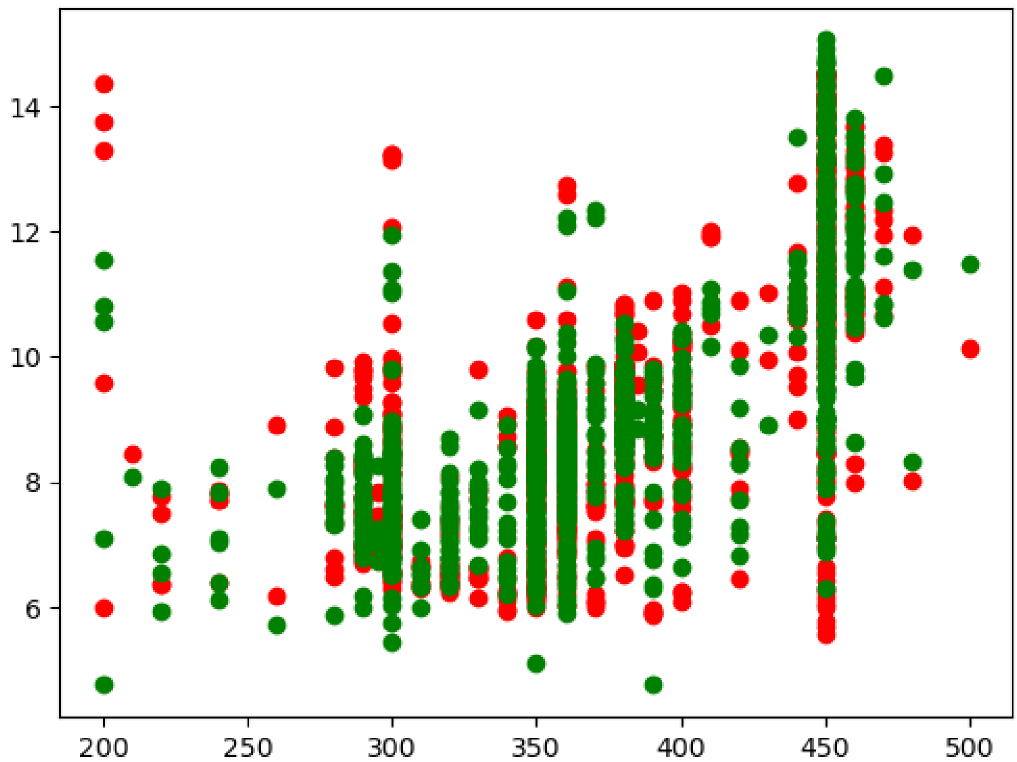
[53]:
```
error=Y-y_predicted
plt.scatter(X['c161'],error)
```

[53]: <matplotlib.collections.PathCollection at 0x2a4c6469780>



[55]:
```
#I have considered to plot the errors against the column 161 as it was having p␣
↪value significantly less than 0.05
#and also the minimum stderror
plt.scatter(X['c161'],Y,color='Red')
plt.scatter(X['c161'],y_predicted,color='Green')
```

[55]: <matplotlib.collections.PathCollection at 0x2a4c648b790>



[70]:
```
#The above is the predicted Data before removing the least significant columns

# The columns having a p-value of greater than 0.05 are:
# c1,c2,c27,c32,c33,c160,c162,c7,c12,c13,c19,c34,c22,c36
# Out of which c36 is having the highest p-value of  0.972

# So,now we are going to remove those columns from X and perform the analysis␣
↪again!

X_final_1=X.
↪drop(['c1','c27','c32','c33','c160','c162','c7','c12','c13','c19','c34','c22','c36'],axis=1
```

[71]:
```
mlr_model_final_1= sm.OLS(Y, X_final_1).fit()
print(mlr_model_final_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    c52   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.780
Method:                 Least Squares   F-statistic:                     141.0
```

Date:	Sat, 02 Sep 2023	Prob (F-statistic):	2.60e-312
Time:	22:50:40	Log-Likelihood:	-1476.6
No. Observations:	1025	AIC:	3007.
Df Residuals:	998	BIC:	3140.
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
c2	-72.7881	8.132	-8.951	0.000	-88.746	-56.830
c26	0.3867	0.047	8.286	0.000	0.295	0.478
c28	0.1943	0.043	4.540	0.000	0.110	0.278
c29	-0.4637	0.047	-9.896	0.000	-0.556	-0.372
c30	3.4448	0.445	7.735	0.000	2.571	4.319
c31	0.2500	0.031	8.115	0.000	0.190	0.310
c39	18.2020	1.398	13.016	0.000	15.458	20.946
c139	-0.4462	0.041	-10.805	0.000	-0.527	-0.365
c142	-0.2304	0.074	-3.093	0.002	-0.377	-0.084
c143	-0.2376	0.037	-6.439	0.000	-0.310	-0.165
c155	-0.0544	0.011	-4.812	0.000	-0.077	-0.032
c157	0.2468	0.036	6.794	0.000	0.176	0.318
c158	0.2966	0.023	13.137	0.000	0.252	0.341
c161	0.0125	0.001	12.820	0.000	0.011	0.014
c163	0.0085	0.002	4.100	0.000	0.004	0.013
c8	-0.6010	0.130	-4.624	0.000	-0.856	-0.346
c9	-0.7825	0.066	-11.828	0.000	-0.912	-0.653
c10	9.5455	1.451	6.578	0.000	6.698	12.393
c11	0.3271	0.078	4.213	0.000	0.175	0.480
c15	-0.3765	0.051	-7.399	0.000	-0.476	-0.277
c16	-0.5418	0.081	-6.724	0.000	-0.700	-0.384
c17	-0.0549	0.020	-2.728	0.006	-0.094	-0.015
c20	0.1735	0.039	4.502	0.000	0.098	0.249
c21	-0.2025	0.047	-4.329	0.000	-0.294	-0.111
c23	-0.3358	0.043	-7.846	0.000	-0.420	-0.252
c35	8.0013	1.505	5.318	0.000	5.049	10.954
c241	14.2471	1.860	7.660	0.000	10.597	17.897

Omnibus:	47.822	Durbin-Watson:	0.570
Prob(Omnibus):	0.000	Jarque-Bera (JB):	148.804
Skew:	-0.072	Prob(JB):	4.87e-33
Kurtosis:	4.861	Cond. No.	1.78e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.78e+05. This might indicate that there are strong multicollinearity or other numerical problems.

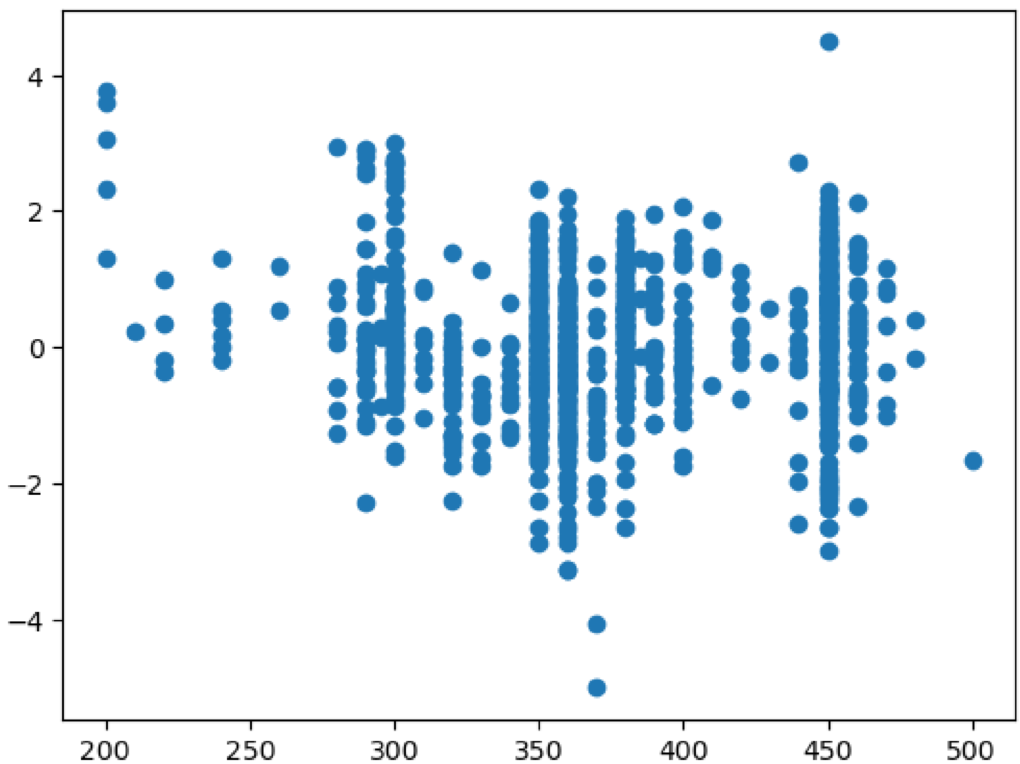
[72]:
```python
#Again we are having columns with p-value greater than 0.05,lets remove it!
# c11,c142,c26
X_final_2=X_final_1.drop(['c11','c142','c26'],axis=1)
mlr_model_final= sm.OLS(Y, X_final_2).fit()
print(mlr_model_final.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    c52   R-squared:                       0.765
Model:                            OLS   Adj. R-squared:                  0.759
Method:                 Least Squares   F-statistic:                     141.5
Date:                Sat, 02 Sep 2023   Prob (F-statistic):          3.12e-295
Time:                        22:50:50   Log-Likelihood:                -1524.9
No. Observations:                1025   AIC:                             3098.
Df Residuals:                    1001   BIC:                             3216.
Df Model:                          23
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c2            -1.5721      3.740     -0.420      0.674      -8.911       5.767
c28            0.1657      0.034      4.855      0.000       0.099       0.233
c29           -0.1243      0.020     -6.339      0.000      -0.163      -0.086
c30            1.3746      0.373      3.681      0.000       0.642       2.107
c31            0.2314      0.020     11.352      0.000       0.191       0.271
c39           16.4162      1.430     11.482      0.000      13.611      19.222
c139          -0.4772      0.041    -11.544      0.000      -0.558      -0.396
c143          -0.1499      0.037     -4.022      0.000      -0.223      -0.077
c155          -0.0609      0.012     -5.293      0.000      -0.084      -0.038
c157           0.2561      0.038      6.739      0.000       0.181       0.331
c158           0.3314      0.023     14.339      0.000       0.286       0.377
c161           0.0126      0.001     12.493      0.000       0.011       0.015
c163           0.0104      0.002      4.845      0.000       0.006       0.015
c8            -0.5452      0.135     -4.040      0.000      -0.810      -0.280
c9            -0.7618      0.068    -11.200      0.000      -0.895      -0.628
c10           10.8079      1.501      7.199      0.000       7.862      13.754
c15           -0.4459      0.053     -8.486      0.000      -0.549      -0.343
c16           -0.4630      0.080     -5.797      0.000      -0.620      -0.306
c17           -0.0489      0.020     -2.386      0.017      -0.089      -0.009
c20            0.2192      0.040      5.546      0.000       0.142       0.297
c21           -0.1714      0.048     -3.558      0.000      -0.266      -0.077
c23           -0.3257      0.042     -7.669      0.000      -0.409      -0.242
c35            8.2017      1.567      5.234      0.000       5.127      11.277
c241           5.1962      0.965      5.384      0.000       3.302       7.090
==============================================================================
Omnibus:                       23.778   Durbin-Watson:                   0.512
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49.091
```

```
Skew:                          -0.035   Prob(JB):                     2.19e-11
Kurtosis:                       4.070   Cond. No.                     5.60e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 5.6e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

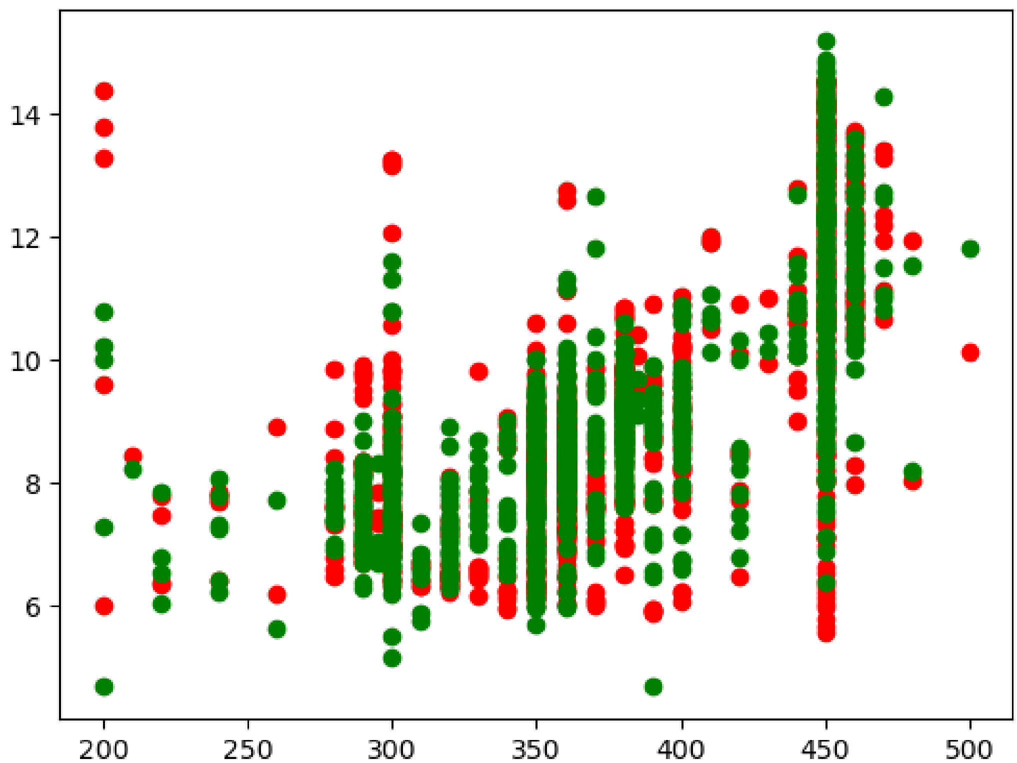
[73]:
```
y_predicted_final=mlr_model_final.predict(X_final_2)
y_predicted_final.head()

error_final=Y-y_predicted_final
plt.scatter(X_final_2['c161'],error_final)
```

[73]: <matplotlib.collections.PathCollection at 0x2a4c65655a0>



[74]:
```
plt.scatter(X_final_2['c161'],Y,color='Red')
plt.scatter(X_final_2['c161'],y_predicted_final,color='Green')
```

[74]: <matplotlib.collections.PathCollection at 0x2a4c6600f70>



[75]:
```
#The above is a better model with all the columns having p-value <0.05 which␣
↪are significantly impacting the given model
#The above regression model is ceentered on c2,now if we remove c2,then we will␣
↪have a better value of R2
```

[81]:
```
X_final_3=X.
↪drop(['c1','c27','c12','c2','c32','c33','c160','c162','c7','c12','c13','c19','c34','c22','c
X_final_4=X_final_3.drop(['c11','c142','c26'],axis=1)
```

[82]:
```
mlr_model_final= sm.OLS(Y, X_final_4).fit()
print(mlr_model_final.summary())
```

```
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                    c52   R-squared (uncentered):
0.987
Model:                            OLS   Adj. R-squared (uncentered):
0.987
Method:                 Least Squares   F-statistic:
3287.
Date:                Sat, 02 Sep 2023   Prob (F-statistic):
```

0.00
Time: 22:53:59 Log-Likelihood:
-1525.0
No. Observations: 1025 AIC:
3096.
Df Residuals: 1002 BIC:
3209.
Df Model: 23
Covariance Type: nonrobust

	coef	std err	t	P>|t|	[0.025	0.975]
c28	0.1673	0.034	4.932	0.000	0.101	0.234
c29	-0.1286	0.017	-7.726	0.000	-0.161	-0.096
c30	1.3160	0.346	3.800	0.000	0.636	1.996
c31	0.2321	0.020	11.427	0.000	0.192	0.272
c39	16.2452	1.370	11.858	0.000	13.557	18.934
c139	-0.4766	0.041	-11.541	0.000	-0.558	-0.396
c143	-0.1499	0.037	-4.023	0.000	-0.223	-0.077
c155	-0.0602	0.011	-5.289	0.000	-0.083	-0.038
c157	0.2519	0.037	6.869	0.000	0.180	0.324
c158	0.3302	0.023	14.398	0.000	0.285	0.375
c161	0.0126	0.001	12.495	0.000	0.011	0.015
c163	0.0103	0.002	4.837	0.000	0.006	0.014
c8	-0.5489	0.135	-4.078	0.000	-0.813	-0.285
c9	-0.7603	0.068	-11.198	0.000	-0.894	-0.627
c10	10.7765	1.499	7.190	0.000	7.835	13.718
c15	-0.4454	0.053	-8.482	0.000	-0.548	-0.342
c16	-0.4698	0.078	-6.008	0.000	-0.623	-0.316
c17	-0.0491	0.020	-2.399	0.017	-0.089	-0.009
c20	0.2180	0.039	5.532	0.000	0.141	0.295
c21	-0.1746	0.048	-3.675	0.000	-0.268	-0.081
c23	-0.3295	0.041	-7.952	0.000	-0.411	-0.248
c35	8.0187	1.505	5.329	0.000	5.066	10.971
c241	5.0796	0.924	5.497	0.000	3.266	6.893

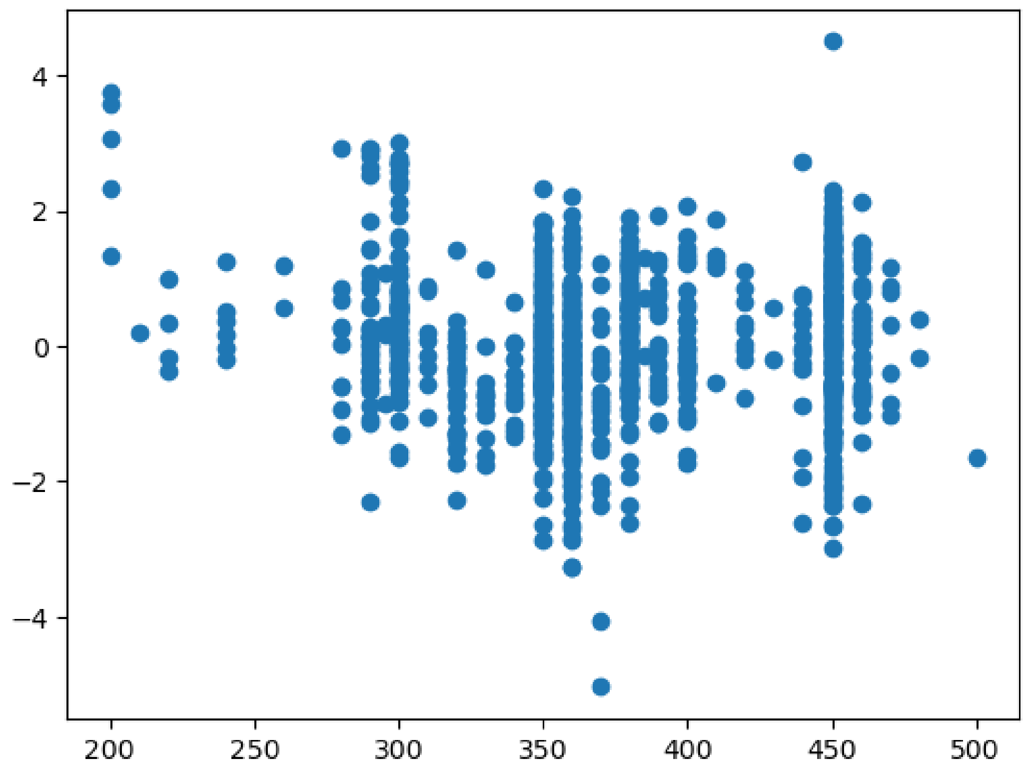
Omnibus:	23.936	Durbin-Watson:	0.510
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49.404
Skew:	-0.041	Prob(JB):	1.87e-11
Kurtosis:	4.072	Cond. No.	2.73e+04

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 2.73e+04. This might indicate that there are

strong multicollinearity or other numerical problems.

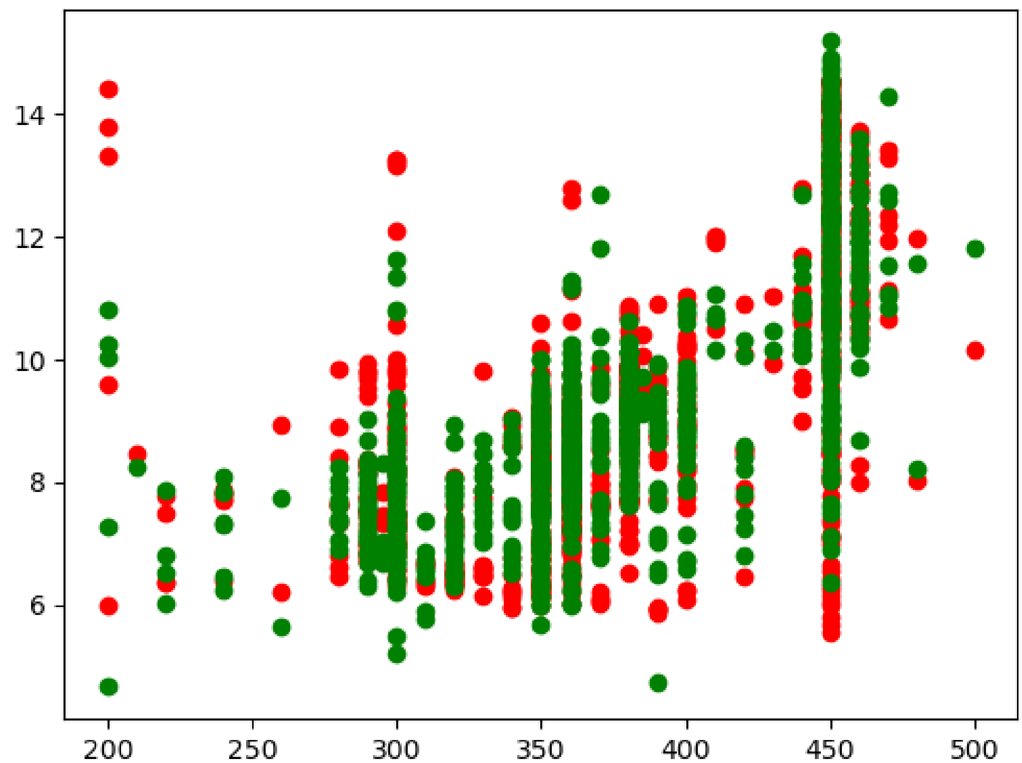
[84]:
```
#Now we can see that the value of R2 has increases considerably from 0.765 to 0.
↪987!!
#This occurs when we remove the column c2!
y_predicted_final_4=mlr_model_final.predict(X_final_4)
y_predicted_final_4.head()

error_final_4=Y-y_predicted_final_4
plt.scatter(X_final_4['c161'],error_final_4)
```

[84]: <matplotlib.collections.PathCollection at 0x2a4c67d5ea0>



[85]:
```
plt.scatter(X_final_4['c161'],Y,color='Red')
plt.scatter(X_final_4['c161'],y_predicted_final_4,color='Green')
```

[85]: <matplotlib.collections.PathCollection at 0x2a4c6877ca0>



In my analysis the most significant variable is the column number c39 because it is having a larger magnitude for the coefficient and also a smaller std error