Assignment 4

This assignment asks you to use resources at hand to apply module 6 - Linear Regression to several sets of data.

Learning Outcomes

- Exploratory analysis for regression
- Understand difference between linear and non-linear models
- Carry out OLS regression model
- Evaluate model

Question 1

- For each data set in Assignment4_linear_regression_data.xlsx:
- Create a scatter plot and visually decide if a linear model is appropriate (a matrix scatter plot will would be most efficient).
- If the relation is not linear, transform the data accordingly.
    - Try logarithm, exponential, square root, square, etc., for Y and/or X until you see a linear relation. You only need to report what is the transformation chosen, not all the attempts. Note: most of the time, you can guess visually. A systematic way is to create a matrix scatter plot of the different transformations. A generic way we did not cover is to use a Box-Cox transformation.
- Create an OLS model for the original and transformed data if required.
    - Evaluate if the OLS assumptions are met: normality of errors centered around zero, equal variance, etc..., for the original data and transformed data if appropriate.
    - Comment how the transformation impacted the different assumptions. (This should be done only by looking at the output diagnostic charts created by the software)
    - If datasets have outliers, remove the outliers and see the effect in the model (slope, intercept and R-square)

The output of the assignment should be:

- OLS full report for the original and transformed data if appropriate (only two datasets should need transformation).
- A short comment on the validity of the linear assumptions for the original and transformed data set when appropriate (it should not need to be longer than a couple of sentences).
- An interpretation of the slope and intercept in relation to the original data, i.e. if the model is linear [intercept value] is the expected value when the independent variable is zero, etc.). If the model is not linear, you need to transform the equation back to its original form.

Check out the following if you need further guidance:
http://www.bzst.com/2009/09/interpreting-log-transformed-variables.html

https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/

https://stats.idre.ucla.edu/sas/faq/how-can-i-interpret-log-transformed-variables-in-terms-of-percent-change-in-linear-regression/

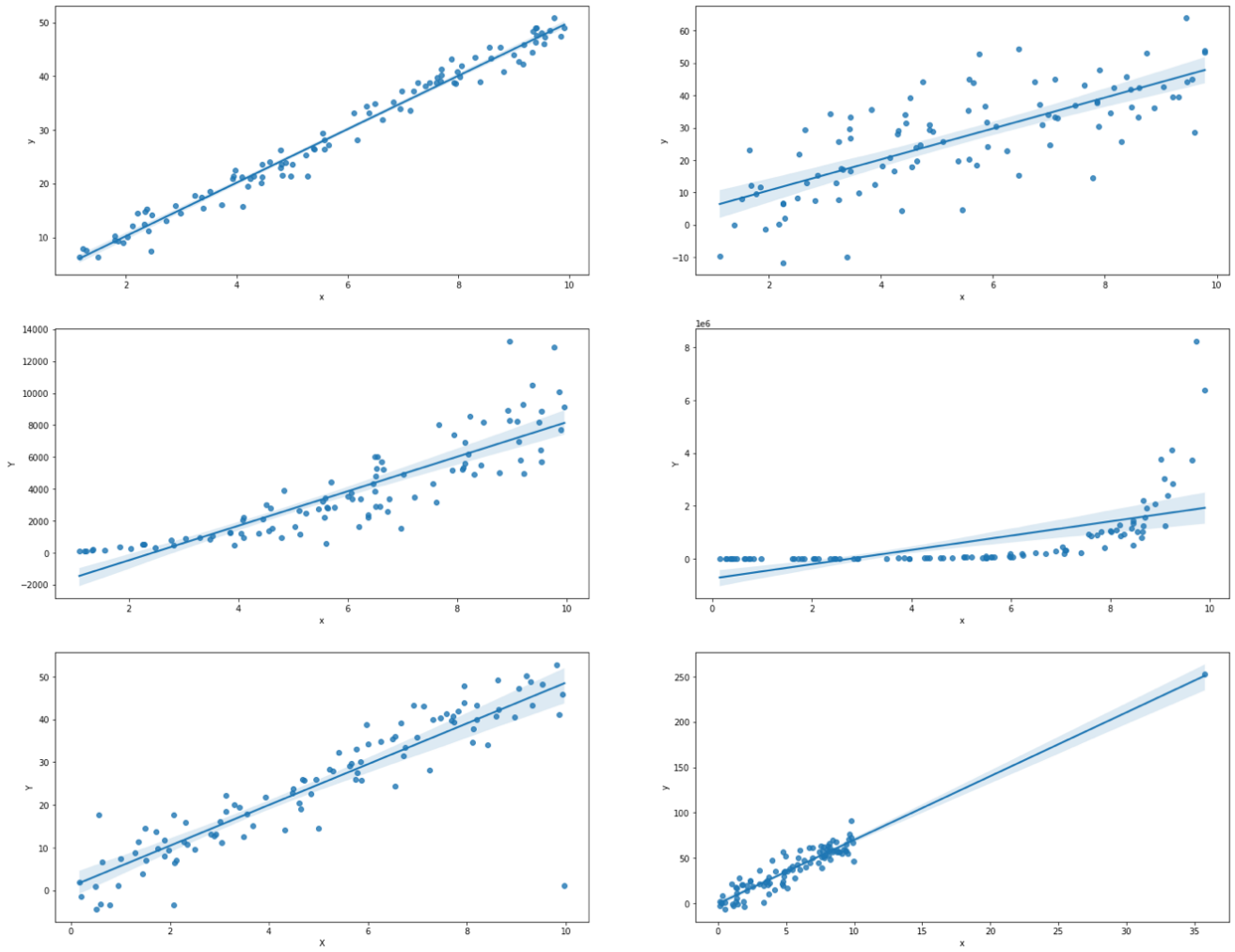
https://stats.stackexchange.com/questions/266722/interpretation-of-linear-regression-results-where-dependent-variable-is-transfor

- If the dataset have outliers, determine if the outlier have leverage or not by comparing the OLS with and without the outlier.

In [1]:
```
import pandas as pd
import openpyxl
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import copy
from scipy import stats
import statsmodels.api as sm
import statsmodels.formula.api as smf
import math

NUMBEROFSETS = 6
FILENAME = 'Assignment4_linear_regresion_data.xlsx'
Y = 0
X = 1
INTERCEPT = 0
SLOPE = 1
```

In [2]:
```
dfs = []
for i in range(1, NUMBEROFSETS+1):
    dfs.append(pd.read_excel(FILENAME, sheet_name='Set '+str(i)))
```

In [3]:
```
figure, axis = plt.subplots(3, 2, figsize=(25, 20))

for i in range(0, 3):
    for j in range(0, 2):
        sns.regplot(y=dfs[2*i+j].columns[Y], x= dfs[2*i+j].columns[X], data=d
plt.show()
```



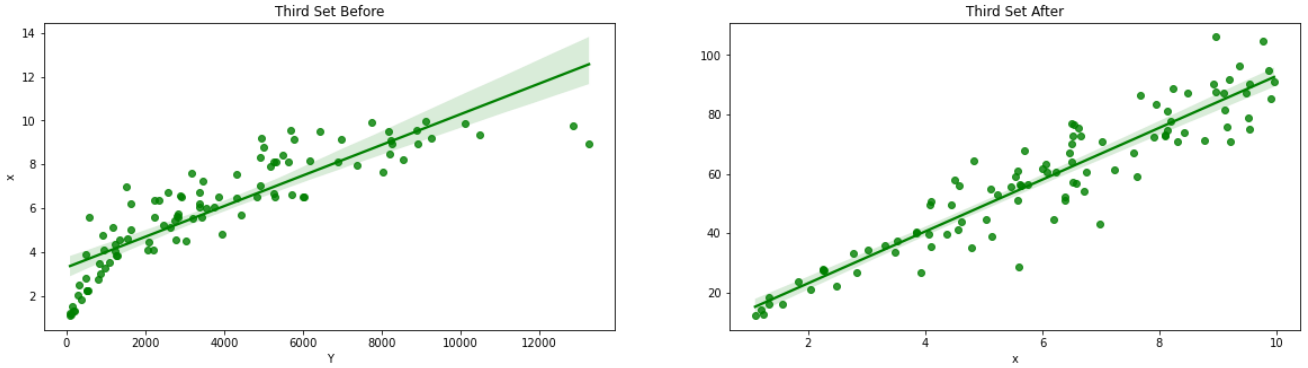
All the data sets are linear except third and 4th data sets.

In [4]:
```
linear_dfs = copy.deepcopy(dfs)
```

For the third plot we use Box-Cox transformation

In [5]:
```
figure, axis = plt.subplots(1, 2, figsize=(20, 5))
sns.regplot(y=dfs[2].columns[X], x= dfs[2].columns[Y], data=dfs[2], ax=axis[0
axis[0].set_title("Third Set Before")
sns.regplot(x=dfs[2].columns[X], y=stats.boxcox(dfs[2][dfs[2].columns[Y]])[0]
            ax=axis[1], color='green')
axis[1].set_title("Third Set After")
```

Out[5]:
```
Text(0.5, 1.0, 'Third Set After')
```

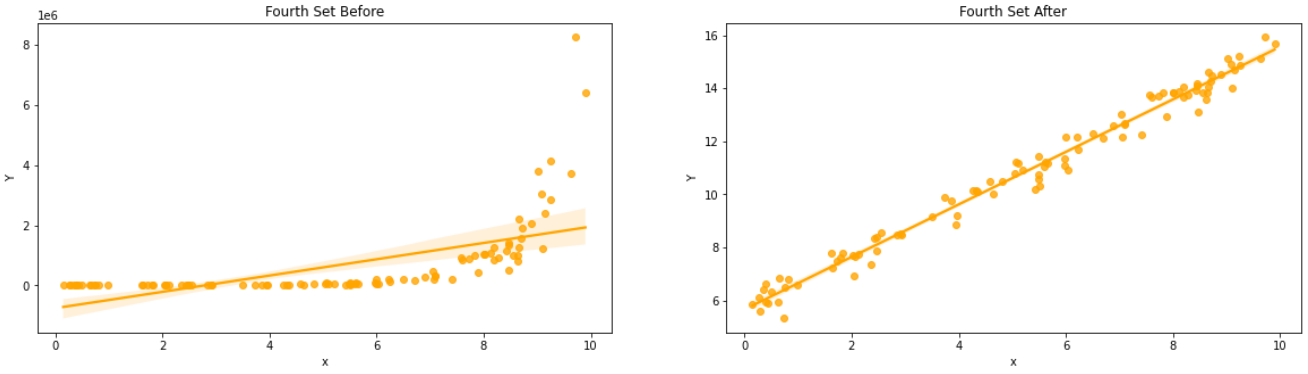


In [6]:
```
linear_dfs[2][linear_dfs[2].columns[Y]] = stats.boxcox(dfs[2][dfs[2].columns[
```

The fourth plot is an exponentioal plot, so to transform it, we should apply log function on y.

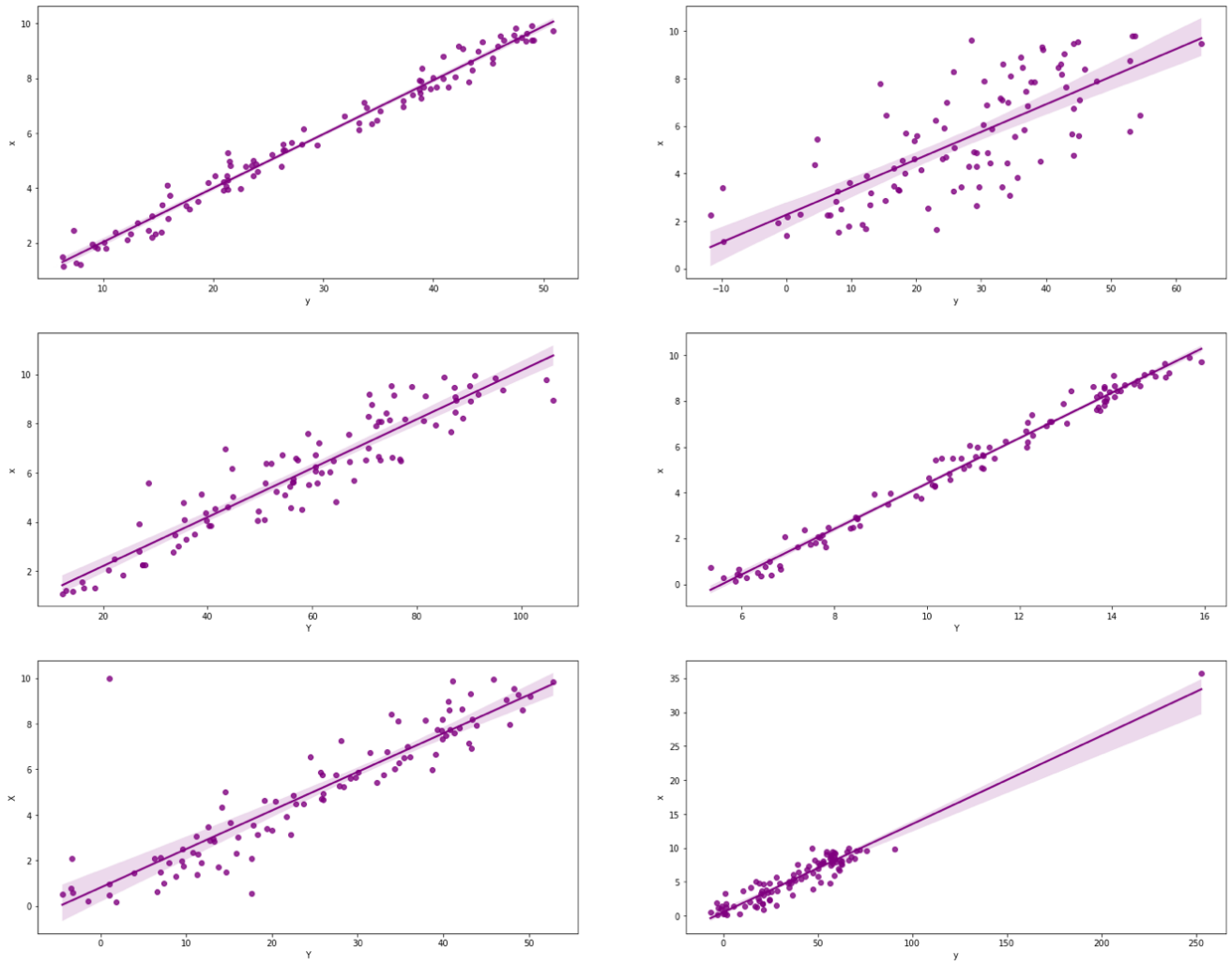
In [7]:
```
figure, axis = plt.subplots(1, 2, figsize=(20, 5))
sns.regplot(y=dfs[3].columns[Y], x= dfs[3].columns[X], data=dfs[3], ax=axis[0
axis[0].set_title("Fourth Set Before")
sns.regplot(x=dfs[3].columns[X], y=np.log(dfs[3][dfs[3].columns[Y]]), data=df
            ax=axis[1], color='orange')
axis[1].set_title("Fourth Set After")
```

Out[7]:
```
Text(0.5, 1.0, 'Fourth Set After')
```



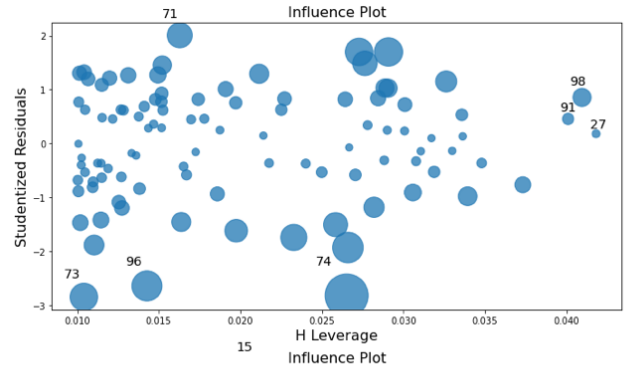
In [8]:
```
linear_dfs[3][linear_dfs[3].columns[Y]] = np.log(dfs[3][dfs[3].columns[Y]])
```

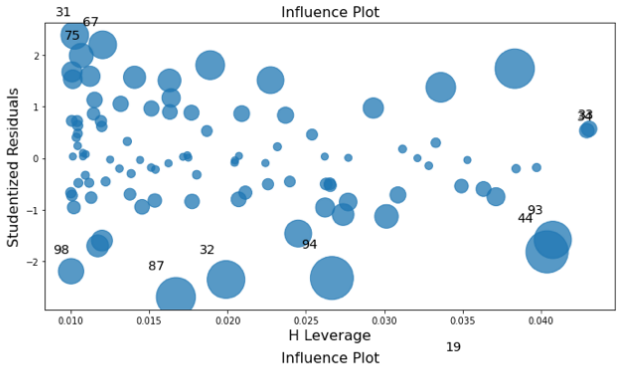
In [9]:
```
figure, axis = plt.subplots(3, 2, figsize=(25, 20))

for i in range(0, 3):
    for j in range(0, 2):
        sns.regplot(y=linear_dfs[2*i+j].columns[X], x= linear_dfs[2*i+j].colu
                    data=linear_dfs[2*i+j], ax=axis[i, j], color='purple')
plt.show()
```

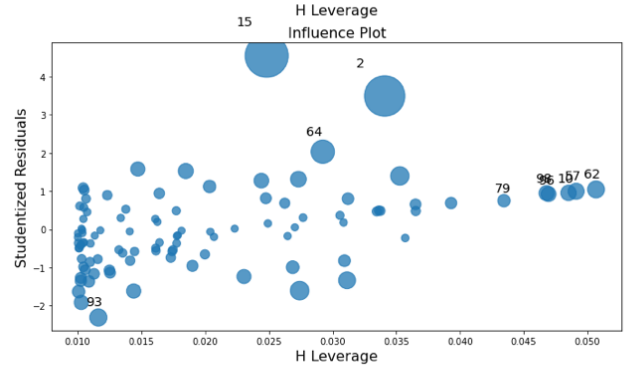


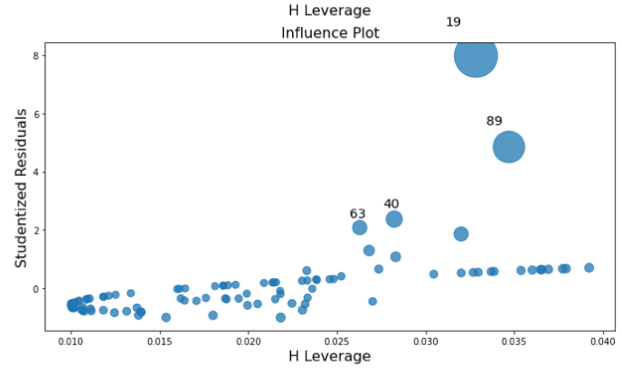
The following is the influence plot for the original data.

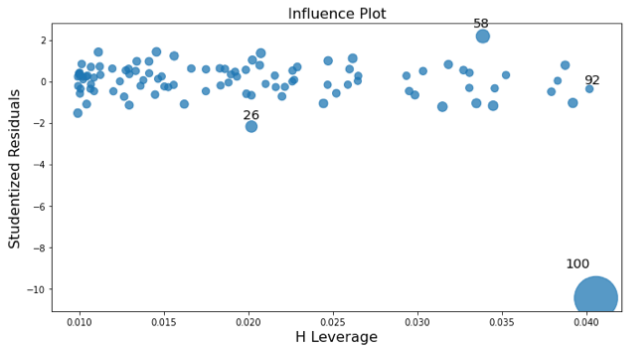
In [10]:
```
ols_models = []
figure, axis = plt.subplots(3, 2, figsize=(25, 20))

for i in range(0, 3):
    for j in range(0, 2):
        results = smf.ols(dfs[2*i+j].columns[Y]+' ~ '+dfs[2*i+j].columns[X], 
        ols_models.append(results)
        sm.graphics.influence_plot(results, ax=axis[i, j])

plt.show()
```

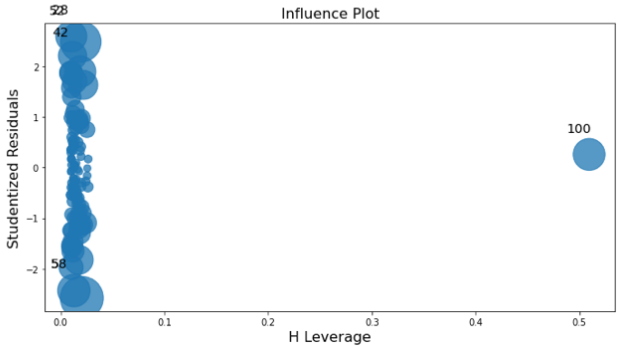












OLS model summary for original data

In [11]:
```
for i in range(0, 6):
    print("-------------------------------    SET " + str(i+1) + "    ---------
    print(ols_models[i].summary())
    print("\n\n")
```

```
-------------------------------    SET 1    ---------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.979
Model:                            OLS   Adj. R-squared:                  0.979
Method:                 Least Squares   F-statistic:                     4579.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           4.47e-84
Time:                        02:24:50   Log-Likelihood:                -206.03
No. Observations:                 100   AIC:                             416.1
Df Residuals:                      98   BIC:                             421.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

Intercept	0.2381	0.469	0.508	0.613	-0.693	1.169
x	4.9843	0.074	67.669	0.000	4.838	5.130

Omnibus:	4.971	Durbin-Watson:	1.982
Prob(Omnibus):	0.083	Jarque-Bera (JB):	4.783
Skew:	-0.536	Prob(JB):	0.0915
Kurtosis:	2.988	Cond. No.	15.9

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

---------- SET 2 ----------

OLS Regression Results

Dep. Variable:	y	R-squared:	0.555
Model:	OLS	Adj. R-squared:	0.551
Method:	Least Squares	F-statistic:	122.4
Date:	Mon, 29 Nov 2021	Prob (F-statistic):	6.11e-19
Time:	02:24:50	Log-Likelihood:	-375.73
No. Observations:	100	AIC:	755.5
Df Residuals:	98	BIC:	760.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.0956	2.547	0.430	0.668	-3.958	6.149
x	4.7774	0.432	11.062	0.000	3.920	5.634

Omnibus:	0.254	Durbin-Watson:	2.043
Prob(Omnibus):	0.881	Jarque-Bera (JB):	0.079
Skew:	-0.065	Prob(JB):	0.961
Kurtosis:	3.045	Cond. No.	14.7

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

---------- SET 3 ----------

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.755
Model:	OLS	Adj. R-squared:	0.753
Method:	Least Squares	F-statistic:	302.4
Date:	Mon, 29 Nov 2021	Prob (F-statistic):	1.04e-31

Time:	02:24:50	Log-Likelihood:	-873.07
No. Observations:	100	AIC:	1750.
Df Residuals:	98	BIC:	1755.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2636.1748	402.741	-6.546	0.000	-3435.400	-1836.949
x	1081.8266	62.216	17.388	0.000	958.361	1205.292

Omnibus:	21.170	Durbin-Watson:	2.159
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37.896
Skew:	0.863	Prob(JB):	5.90e-09
Kurtosis:	5.474	Cond. No.	17.6

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

---------- SET 4 ----------

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.380
Model:	OLS	Adj. R-squared:	0.373
Method:	Least Squares	F-statistic:	59.97
Date:	Mon, 29 Nov 2021	Prob (F-statistic):	8.87e-12
Time:	02:24:50	Log-Likelihood:	-1526.2
No. Observations:	100	AIC:	3056.
Df Residuals:	98	BIC:	3062.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.535e+05	2.1e+05	-3.585	0.001	-1.17e+06	-3.36e+05
x	2.707e+05	3.49e+04	7.744	0.000	2.01e+05	3.4e+05

Omnibus:	102.143	Durbin-Watson:	2.077
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1253.666
Skew:	3.381	Prob(JB):	5.89e-273
Kurtosis:	18.973	Cond. No.	12.4

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

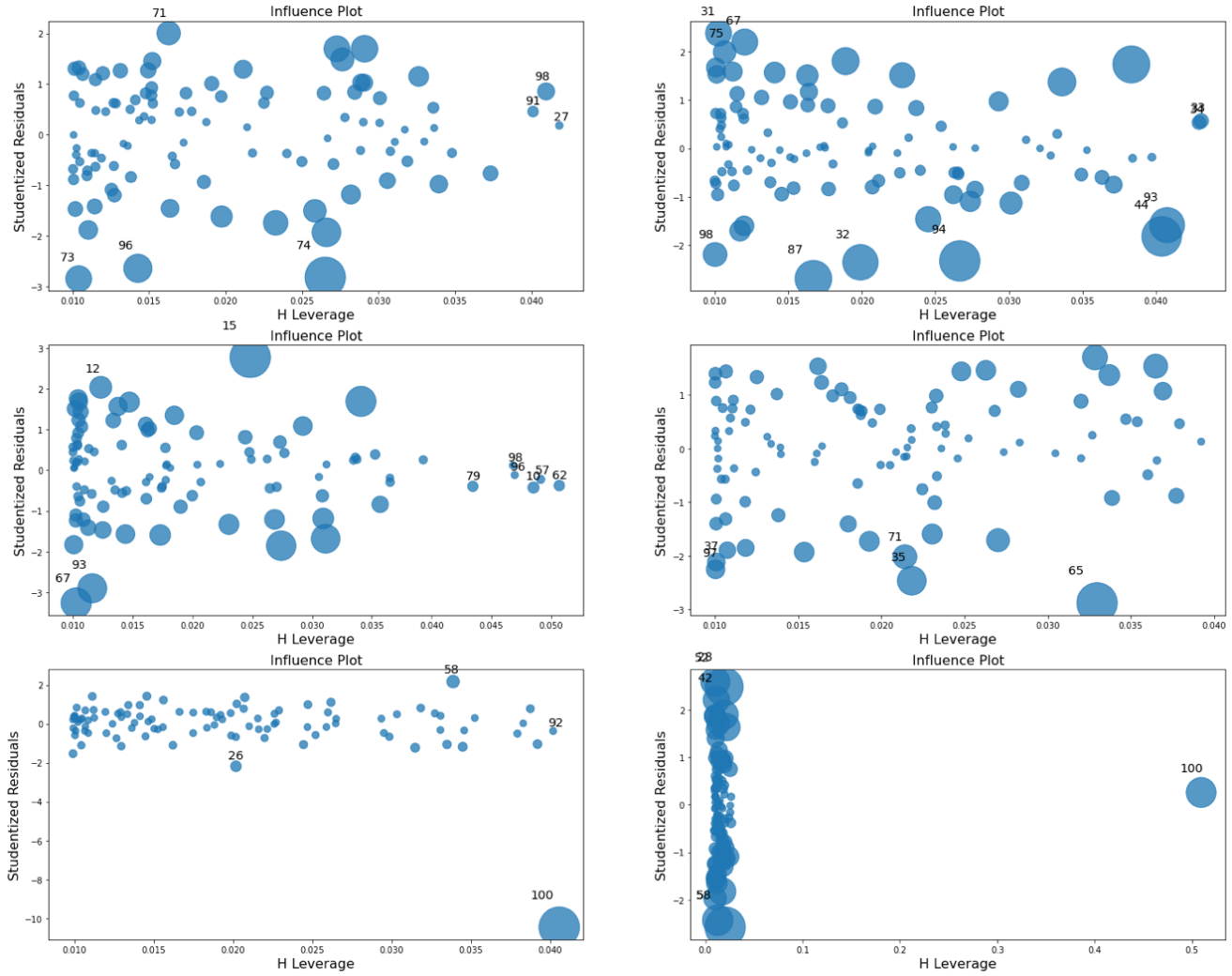
```
--------------------------------   SET 5   ------------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.806
Model:                            OLS   Adj. R-squared:                  0.804
Method:                 Least Squares   F-statistic:                     411.9
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           4.70e-37
Time:                        02:24:50   Log-Likelihood:                -334.42
No. Observations:                 101   AIC:                             672.8
Df Residuals:                      99   BIC:                             678.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.9213      1.346      0.685      0.495      -1.749       3.591
X              4.7671      0.235     20.294      0.000       4.301       5.233
==============================================================================
Omnibus:                      113.783   Durbin-Watson:                   1.491
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2578.951
Skew:                          -3.591   Prob(JB):                         0.00
Kurtosis:                      26.691   Cond. No.                         11.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.



--------------------------------   SET 6   ------------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.913
Model:                            OLS   Adj. R-squared:                  0.912
Method:                 Least Squares   F-statistic:                     1041.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           2.49e-54
Time:                        02:24:50   Log-Likelihood:                -367.52
No. Observations:                 101   AIC:                             739.0
Df Residuals:                      99   BIC:                             744.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.3059      1.534     -0.199      0.842      -3.350       2.739
x              7.0272      0.218     32.259      0.000       6.595       7.459
==============================================================================
Omnibus:                        0.494   Durbin-Watson:                   2.255
Prob(Omnibus):                  0.781   Jarque-Bera (JB):                0.262
Skew:                           0.120   Prob(JB):                        0.877
Kurtosis:                       3.070   Cond. No.                         11.8
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

influence plot and summary for transformed data.

In [12]:

```
ols_models_linear = []
figure, axis = plt.subplots(3, 2, figsize=(25, 20))

for i in range(0, 3):
    for j in range(0, 2):
        results = smf.ols(linear_dfs[2*i+j].columns[Y]+' ~ '+linear_dfs[2*i+j
                          data=linear_dfs[2*i+j]).fit()
        ols_models_linear.append(results)
        sm.graphics.influence_plot(results, ax=axis[i, j])

plt.show()
```



In [13]:
```
for i in range(0, 6):
    print("-------------------------------   SET " + str(i+1) + "   ---------
    print(ols_models_linear[i].summary())
    print("\n\n")
```

```
-------------------------------   SET 1   ----------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.979
Model:                            OLS   Adj. R-squared:                  0.979
Method:                 Least Squares   F-statistic:                     4579.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           4.47e-84
Time:                        02:24:51   Log-Likelihood:                -206.03
No. Observations:                 100   AIC:                             416.1
Df Residuals:                      98   BIC:                             421.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.2381      0.469      0.508      0.613      -0.693       1.169
x              4.9843      0.074     67.669      0.000       4.838       5.130
==============================================================================
Omnibus:                        4.971   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.083   Jarque-Bera (JB):                4.783
Skew:                          -0.536   Prob(JB):                       0.0915
Kurtosis:                       2.988   Cond. No.                         15.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.



-------------------------------   SET 2   ----------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.555
Model:                            OLS   Adj. R-squared:                  0.551
Method:                 Least Squares   F-statistic:                     122.4
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           6.11e-19
Time:                        02:24:51   Log-Likelihood:                -375.73
No. Observations:                 100   AIC:                             755.5
Df Residuals:                      98   BIC:                             760.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.0956      2.547      0.430      0.668      -3.958       6.149
x              4.7774      0.432     11.062      0.000       3.920       5.634
```

```
==============================================================================
Omnibus:                        0.254   Durbin-Watson:                   2.043
Prob(Omnibus):                  0.881   Jarque-Bera (JB):                0.079
Skew:                          -0.065   Prob(JB):                        0.961
Kurtosis:                       3.045   Cond. No.                         14.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

```
--------------------------------   SET 3   ---------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.868
Model:                            OLS   Adj. R-squared:                  0.866
Method:                 Least Squares   F-statistic:                     641.8
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           8.32e-45
Time:                        02:24:51   Log-Likelihood:                -353.39
No. Observations:                 100   AIC:                             710.8
Df Residuals:                      98   BIC:                             716.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.7418      2.229      2.576      0.011       1.319      10.165
x              8.7229      0.344     25.333      0.000       8.040       9.406
==============================================================================
Omnibus:                        3.044   Durbin-Watson:                   1.854
Prob(Omnibus):                  0.218   Jarque-Bera (JB):                2.483
Skew:                          -0.254   Prob(JB):                        0.289
Kurtosis:                       3.582   Cond. No.                         17.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

```
--------------------------------   SET 4   ---------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.983
Model:                            OLS   Adj. R-squared:                  0.983
Method:                 Least Squares   F-statistic:                     5765.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           6.91e-89
Time:                        02:24:51   Log-Likelihood:                -46.034
No. Observations:                 100   AIC:                             96.07
Df Residuals:                      98   BIC:                             101.3
```

```
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.6647      0.078     72.264      0.000       5.509       5.820
x              0.9898      0.013     75.930      0.000       0.964       1.016
==============================================================================
Omnibus:                        7.020   Durbin-Watson:                   2.151
Prob(Omnibus):                  0.030   Jarque-Bera (JB):                7.256
Skew:                          -0.657   Prob(JB):                       0.0266
Kurtosis:                       2.872   Cond. No.                         12.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.



--------------------------------   SET 5   ------------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.806
Model:                            OLS   Adj. R-squared:                  0.804
Method:                 Least Squares   F-statistic:                     411.9
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           4.70e-37
Time:                        02:24:51   Log-Likelihood:                -334.42
No. Observations:                 101   AIC:                             672.8
Df Residuals:                      99   BIC:                             678.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.9213      1.346      0.685      0.495      -1.749       3.591
X              4.7671      0.235     20.294      0.000       4.301       5.233
==============================================================================
Omnibus:                      113.783   Durbin-Watson:                   1.491
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2578.951
Skew:                          -3.591   Prob(JB):                         0.00
Kurtosis:                      26.691   Cond. No.                         11.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.



--------------------------------   SET 6   ------------------------------------
                            OLS Regression Results
==============================================================================
```

```
Dep. Variable:                      y   R-squared:                       0.913
Model:                            OLS   Adj. R-squared:                  0.912
Method:                 Least Squares   F-statistic:                     1041.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           2.49e-54
Time:                        02:24:51   Log-Likelihood:                -367.52
No. Observations:                 101   AIC:                             739.0
Df Residuals:                      99   BIC:                             744.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.3059      1.534     -0.199      0.842      -3.350       2.739
x              7.0272      0.218     32.259      0.000       6.595       7.459
==============================================================================
Omnibus:                        0.494   Durbin-Watson:                   2.255
Prob(Omnibus):                  0.781   Jarque-Bera (JB):                0.262
Skew:                           0.120   Prob(JB):                        0.877
Kurtosis:                       3.070   Cond. No.                         11.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

In [14]:
```
linear_coefs = [4.9843, 4.7774, 8.7229, 0.9898, 4.7671, 7.0272]
linear_stderrs = [0.074, 0.432, 0.344, 0.013, 0.235, 0.218]
```

influence plot and summary for transformed data when outliners removed.

In [15]:
```
outliners_linear = [[27, 71, 73, 74, 91, 96, 98], [31, 32, 33, 34, 44, 67, 75
             [10, 12, 57, 62, 67, 79, 93, 96, 98], [35, 37, 65, 71, 97], [26,

without_outliners_linear = copy.deepcopy(linear_dfs)
for i in range(0, 6):
    without_outliners_linear[i] = without_outliners_linear[i].drop(outliners_
```

In [16]:
```
without_outliners_ols_models_linear = []
for i in range(0, 6):
    results = smf.ols(without_outliners_linear[i].columns[Y]+' ~ '+without_ou
                      data=without_outliners_linear[i]).fit()
    without_outliners_ols_models_linear.append(results)
```

In [17]:
```
for i in range(0, 6):
    print("------------------------------   SET " + str(i+1) + "   ---------
    print(without_outliners_ols_models_linear[i].summary())
    print("\n\n")
```

```
------------------------------   SET 1   ----------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.982
Model:                            OLS   Adj. R-squared:                  0.982
Method:                 Least Squares   F-statistic:                     5017.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           2.17e-81
Time:                        02:24:52   Log-Likelihood:                -180.30
No. Observations:                  93   AIC:                             364.6
Df Residuals:                      91   BIC:                             369.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.5841      0.454      1.287      0.201      -0.318       1.486
x              4.9422      0.070     70.830      0.000       4.804       5.081
==============================================================================
Omnibus:                        7.468   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.024   Jarque-Bera (JB):                3.483
Skew:                          -0.207   Prob(JB):                        0.175
Kurtosis:                       2.147   Cond. No.                         17.1
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.



------------------------------   SET 2   ----------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.621
Model:                            OLS   Adj. R-squared:                  0.616
Method:                 Least Squares   F-statistic:                     142.3
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           5.29e-20
Time:                        02:24:52   Log-Likelihood:                -314.70
No. Observations:                  89   AIC:                             633.4
Df Residuals:                      87   BIC:                             638.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.0428      2.196      1.385      0.169      -1.323       7.408
x              4.5221      0.379     11.929      0.000       3.769       5.276
```

```
==============================================================================
Omnibus:                        2.193   Durbin-Watson:                   1.954
Prob(Omnibus):                  0.334   Jarque-Bera (JB):                2.036
Skew:                           0.278   Prob(JB):                        0.361
Kurtosis:                       2.511   Cond. No.                         14.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.



--------------------------------   SET 3   ---------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.862
Model:                            OLS   Adj. R-squared:                  0.861
Method:                 Least Squares   F-statistic:                     557.8
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           4.24e-40
Time:                        02:24:52   Log-Likelihood:                -314.20
No. Observations:                  91   AIC:                             632.4
Df Residuals:                      89   BIC:                             637.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.7568      2.448      2.760      0.007       1.892      11.622
x              8.6391      0.366     23.617      0.000       7.912       9.366
==============================================================================
Omnibus:                        0.242   Durbin-Watson:                   2.042
Prob(Omnibus):                  0.886   Jarque-Bera (JB):                0.379
Skew:                           0.106   Prob(JB):                        0.827
Kurtosis:                       2.766   Cond. No.                         20.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.



--------------------------------   SET 4   ---------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.987
Model:                            OLS   Adj. R-squared:                  0.987
Method:                 Least Squares   F-statistic:                     7187.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           7.30e-90
Time:                        02:24:52   Log-Likelihood:                -30.489
No. Observations:                  95   AIC:                             64.98
Df Residuals:                      93   BIC:                             70.09
```

```
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.7360      0.070     81.564      0.000       5.596       5.876
x              0.9851      0.012     84.778      0.000       0.962       1.008
==============================================================================
Omnibus:                        3.725   Durbin-Watson:                   2.233
Prob(Omnibus):                  0.155   Jarque-Bera (JB):                3.709
Skew:                          -0.449   Prob(JB):                        0.157
Kurtosis:                       2.638   Cond. No.                         12.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.



--------------------------------   SET 5   ------------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.921
Model:                            OLS   Adj. R-squared:                  0.920
Method:                 Least Squares   F-statistic:                     1108.
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           3.57e-54
Time:                        02:24:52   Log-Likelihood:                -275.49
No. Observations:                  97   AIC:                             555.0
Df Residuals:                      95   BIC:                             560.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.3862      0.874     -0.442      0.659      -2.121       1.348
X              5.1356      0.154     33.291      0.000       4.829       5.442
==============================================================================
Omnibus:                        2.095   Durbin-Watson:                   2.103
Prob(Omnibus):                  0.351   Jarque-Bera (JB):                2.050
Skew:                          -0.345   Prob(JB):                        0.359
Kurtosis:                       2.821   Cond. No.                         11.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.



--------------------------------   SET 6   ------------------------------------
                            OLS Regression Results
==============================================================================
```

```
Dep. Variable:                      y   R-squared:                       0.850
Model:                            OLS   Adj. R-squared:                  0.848
Method:                 Least Squares   F-statistic:                     542.5
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           2.79e-41
Time:                        02:24:52   Log-Likelihood:                -352.72
No. Observations:                  98   AIC:                             709.4
Df Residuals:                      96   BIC:                             714.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0984      1.835      0.054      0.957      -3.544       3.741
x              6.9521      0.298     23.291      0.000       6.360       7.545
==============================================================================
Omnibus:                        0.764   Durbin-Watson:                   2.186
Prob(Omnibus):                  0.682   Jarque-Bera (JB):                0.523
Skew:                           0.177   Prob(JB):                        0.770
Kurtosis:                       3.053   Cond. No.                         12.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

In [18]:

```
removed_coefs = [4.9422, 4.5221, 8.6391, 0.9851, 5.1356, 6.9521]
removed_stderrs = [0.070, 0.379, 0.366, 0.012, 0.154, 0.298]
```

In [19]:

```
for i in range(0, 6):
    SE = math.sqrt((removed_stderrs[i])**2 + (linear_stderrs[i]**2))
    coef_diff = removed_coefs[i] - linear_coefs[i]
    p_value =  (1 - stats.norm.cdf(coef_diff/SE)) * 2
    print(p_value)
```

```
1.320614361582404
1.3431323136751678
1.1325016669801928
1.2094987022288204
0.1896706582323473
1.1611763032220639
```

H0: the difference in coef values is zero

HA: the difference is non-zero

As you can see p-values for all sets is very high so we cannot reject the H0 so we can conclude that the changes in coef is not significant, so we didn't need to remove the outliners.

In [ ]:

In [ ]: