Task-C: Regression outlier effect.

Objective:Visualization best fit linear regression line for different scenarios

In [5]:

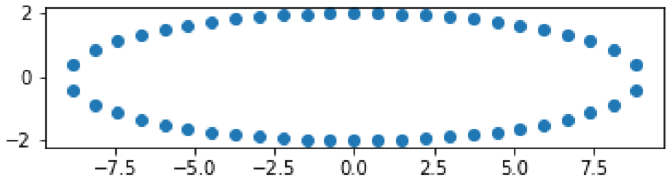
```
# you should not import any other packages
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
import numpy as np
from sklearn.linear_model import SGDRegressor
```

In [2]:

```
import numpy as np
import scipy as sp
import scipy.optimize

def angles_in_ellipse(num,a,b):
    assert(num > 0)
    assert(a < b)
    angles = 2 * np.pi * np.arange(num) / num
    if a != b:
        e = (1.0 - a ** 2.0 / b ** 2.0) ** 0.5
        tot_size = sp.special.ellipeinc(2.0 * np.pi, e)
        arc_size = tot_size / num
        arcs = np.arange(num) * arc_size
        res = sp.optimize.root(
            lambda x: (sp.special.ellipeinc(x, e) - arcs), angles)
        angles = res.x
    return angles
```

In [6]:

```
a = 2
b = 9
n = 50

phi = angles_in_ellipse(n, a, b)
e = (1.0 - a ** 2.0 / b ** 2.0) ** 0.5
arcs = sp.special.ellipeinc(phi, e)

fig = plt.figure()
ax = fig.gca()
ax.axes.set_aspect('equal')
ax.scatter(b * np.sin(phi), a * np.cos(phi))
plt.show()
```



In [38]:

```
X= b * np.sin(phi)
Y= a * np.cos(phi)
print(len(X),len(Y))
```

```
50 50
```

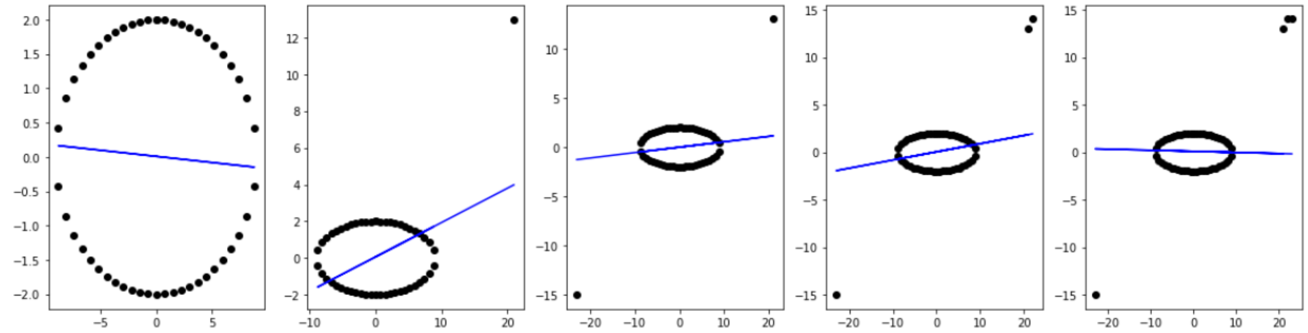
1. As a part of this assignment you will be working the regression problem and how regularization helps to get rid of outliers

2. Use the above created X, Y for this experiment.

3. to do this task you can either implement your own SGDRegression(prefered) excatly similar to "SGD assignment" with mean sequared error or
you can use the SGDRegression of sklearn, for example "SGDRegressor(alpha=0.001, eta0=0.001, learning_rate='constant',random_state=0)"
note that you have to use the constant learning rate and learning rate ***eta0*** *initialized.*

4. as a part of this experiment you will train your linear regression on the data (X, Y) with different regularizations alpha=[0.0001, 1, 100] and
observe how prediction hyper plan moves with respect to the outliers

5. This the results of one of the experiment we did (title of the plot was not metioned intentionally)



in each iteration we were adding single outlier and observed the movement of the hyper plane.

6. please consider this list of outliers: [(0,2),(21, 13), (-23, -15), (22,14), (23, 14)] in each of tuple the first elemet is the input feature(X) and the second element is the output(Y)

7. for each regularizer, you need to add these outliers one at time to data and then train your model
again on the updated data.

*8. you should plot a 3*5 grid of subplots,*
where each row corresponds to results of model with a single regularizer.

9. Algorithm:

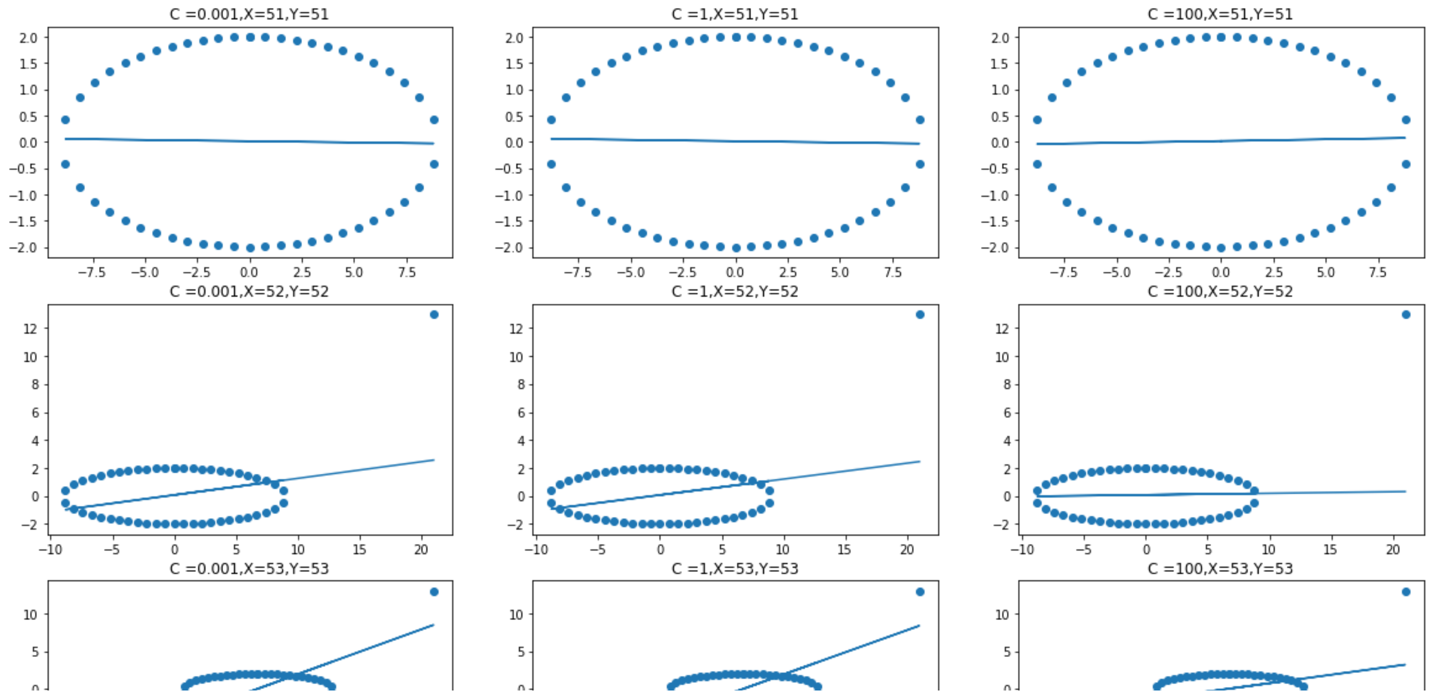
```
for each regularizer:
    for each outlier:
        #add the outlier to the data
        #fit the linear regression to the updated data
        #get the hyper plane
        #plot the hyperplane along with the data points
```

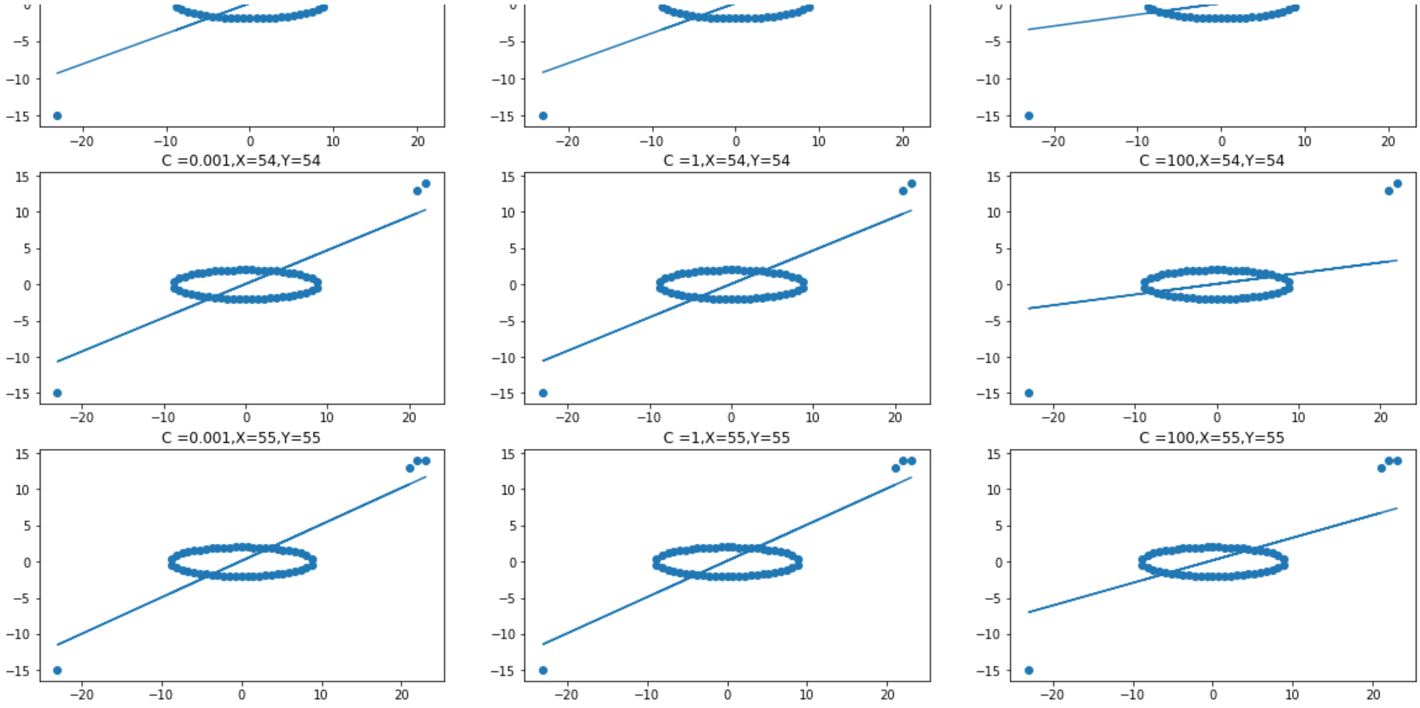
10. MAKE SURE YOU WRITE THE DETAILED OBSERVATIONS, PLEASE CHECK THE LOSS FUNCTION IN THE SKLEARN DOCUMENTATION
(please do search for it).

In [39]:

```
C=[0.001, 1, 100]
k=0
def draw_line(coef,intercept, mi, ma):
    # for the separating hyper plane ax+by+c=0, the weights are [a, b] and the intercept is c
    # to draw the hyper plane we are creating two points
    # 1. ((b*min-c)/a, min) i.e ax+by+c=0 ==> ax = (-by-c) ==> x = (-by-c)/a here in place of y we are keeping the minimum value of y
    # 2. ((b*max-c)/a, max) i.e ax+by+c=0 ==> ax = (-by-c) ==> x = (-by-c)/a here in place of y we are keeping the maximum value of y
    points=np.array([[((-coef[1]*mi - intercept)/coef[0]), mi],[((-coef[1]*ma - intercept)/coef[0]), ma]])
    #points=np.array([[((-coef[0]*mi - intercept)/coef[0]), mi],[((-coef[0]*ma - intercept)/coef[0]), ma]])
    plt.plot(points[:,0], points[:,1])

X=X.reshape(-1,1)
print(len(X))
clf=SGDRegressor(alpha=0.001, eta0=0.001, learning_rate='constant',random_state=0)

clf.fit(X,Y)

print(clf.coef_,clf.intercept_)

outlier=[(0,2),(21, 13), (-23, -15), (22,14), (23, 14)]
#Y=np.append(Y, outlier[0][0])


plt.figure(figsize=(20,20))
for i in outlier:
    X=np.vstack([X, i[0]])
    Y=np.append(Y, i[1])

    for j in C:
        plt.subplot(5, 3, k+1)
        k=k+1
        clf=SGDRegressor(alpha=j, eta0=0.001, learning_rate='constant',random_state=0)
        clf.fit(X,Y)
        y_pred=clf.predict(X)
        plt.title('C ='+str(j)+','+"X="+str(len(X))+','+"Y="+str(len(Y)))
        plt.scatter(X,Y)
        plt.plot(X,y_pred)



plt.show()
```

```
50
[0.01244644] [-0.00171103]
```





C =0.001,X=54,Y=54

C =1,X=54,Y=54

C =100,X=54,Y=54

C =0.001,X=55,Y=55

C =1,X=55,Y=55

C =100,X=55,Y=55

OBSERVATION

1.When C=0.001 ,As the number of outlier increases the regressor line tries to fit it the outlier more, even a small number of outlier makes the model ouput worse .

2.when C=1, as the number of outlier increase ,the line tries to fit it but less as compared when c=0.001 ,here also model can perform bad even with less outlier.

3.when c=100, model is no much impacted by outliers as compared when C is 0.001 and 1 model performs better with low number of outliers.

4.Overall we can conclude that, as C increases the model performs better with less outlier present in data.outlier impact less when c value is high.