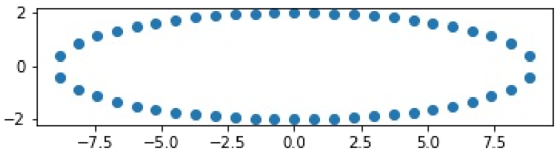
Task-C: Regression outlier effect.

Objective:Visualization best fit linear regression line for different scenarios

In [1]:
```
# you should not import any other packages
import matplotlib.pyplot as plt
import numpy as np
from sklearn.linear_model import SGDRegressor

import warnings
warnings.filterwarnings("ignore")
```

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 [3]:
```
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 [4]:
```
X = b * np.sin(phi)
Y = a * np.cos(phi)
```

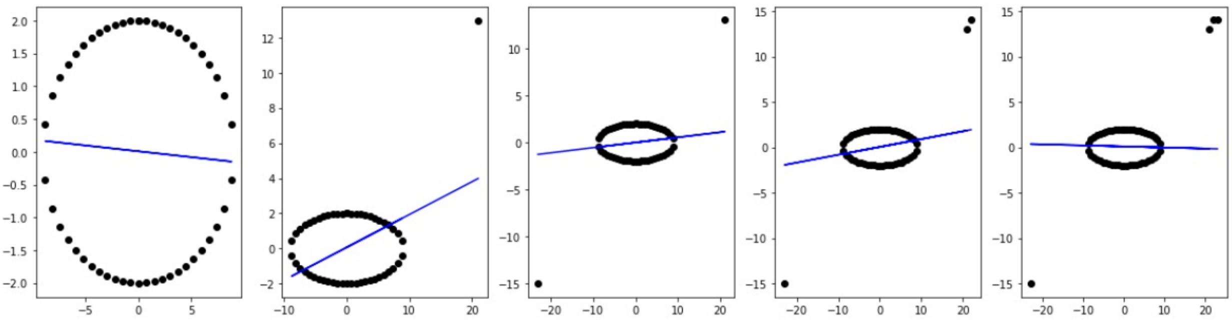
1. As a part of this assignment you will be working the regression problem and how regularization helps to get rid of outliers

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.

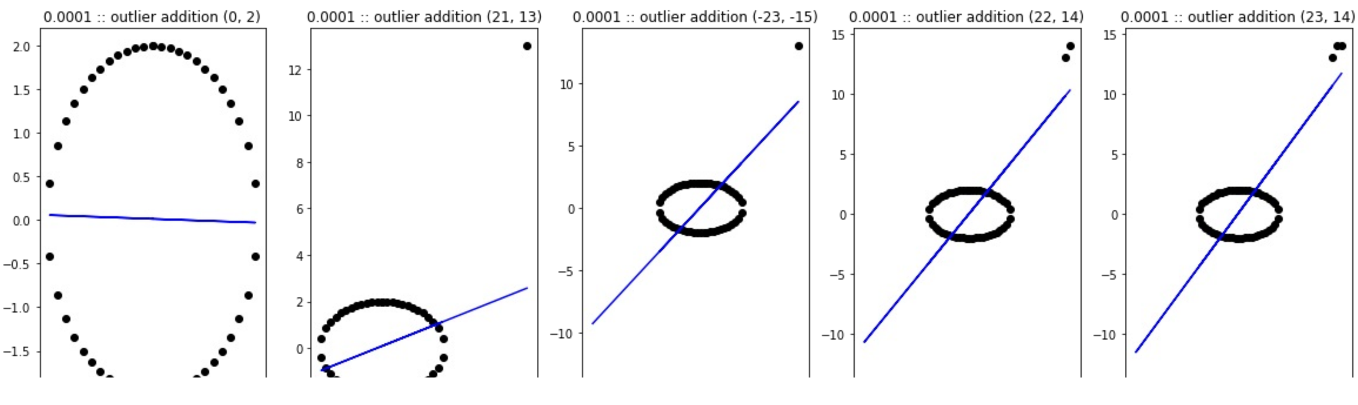
9. Algorithm:

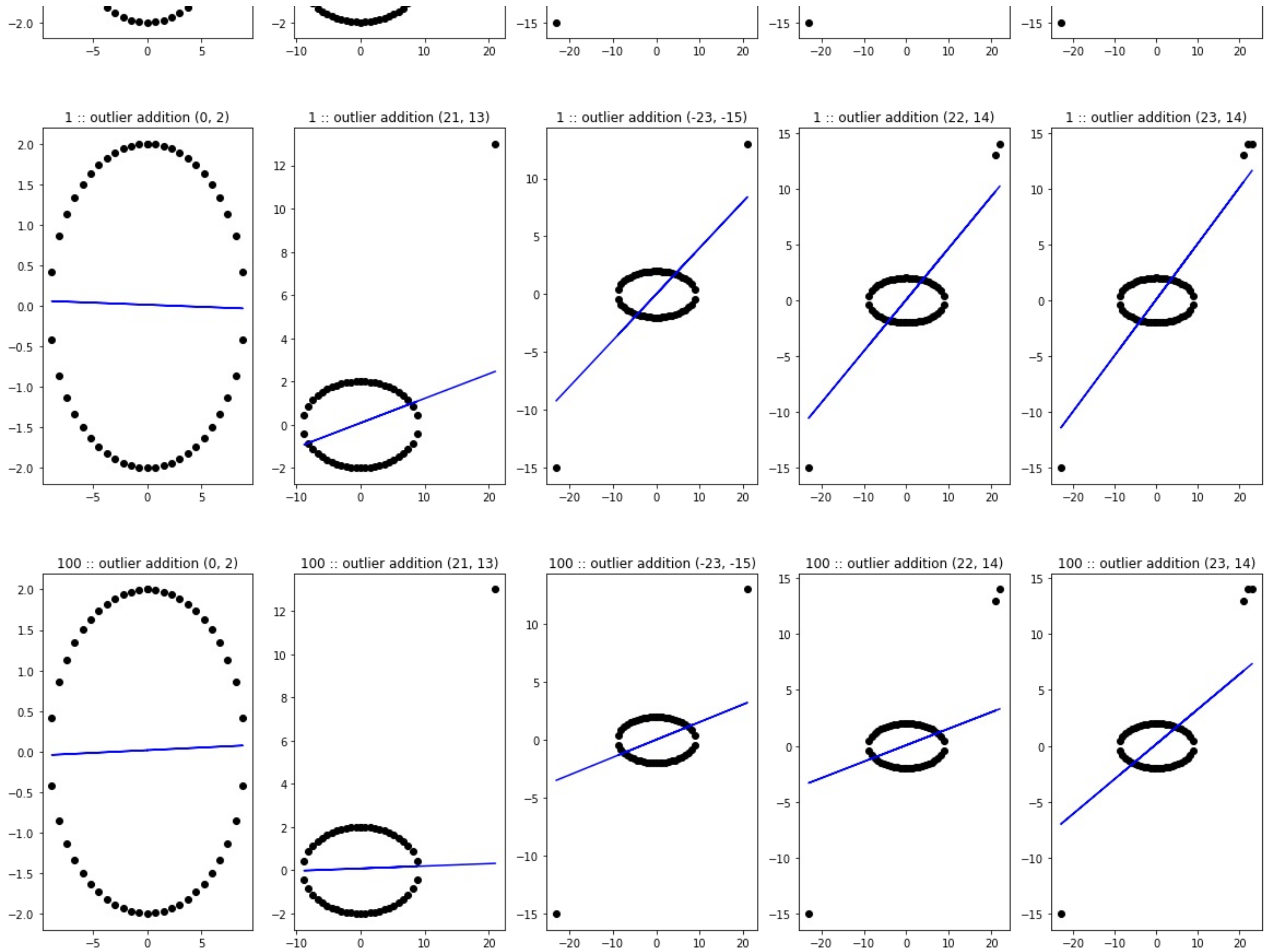
for each regularizer:
  for each outlier:
    #add the outlier to the data
    #fit the linear regression to the updated data
    #get the hyper plane
    #plot the hyperplane along with the data points

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

In [5]:

```
# https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html

alphas = [0.0001, 1, 100]
outliers = [(0,2),(21, 13), (-23, -15), (22,14), (23, 14)]

# https://stackoverflow.com/a/17211410
fig, axs  = plt.subplots(3, 5, figsize=(20, 21))
fig.subplots_adjust(hspace = 0.25)
axs = axs.ravel()
i = 0

for alpha in alphas:
    for outlier in outliers:
        X = np.append(X, outlier[0])
        Y = np.append(Y, outlier[1])
        sgd_reg = SGDRegressor(alpha = alpha, eta0 = 0.001, learning_rate = 'constant',random_state = 0)
        sgd_reg.fit(X.reshape(-1, 1), Y)
        Y_pred = sgd_reg.predict(X.reshape(-1, 1))

        '''https://gist.githubusercontent.com/adarsh1021/9a9b63669bfd338d875307dbfc71de06/
                    raw/9ef0dbde689125977533fa603cf202bf07395834/visualize.py'''

        axs[i].scatter(X, Y, color = 'black')
        axs[i].plot(X, Y_pred, color='blue')
        axs[i].set_title(f'{alpha} :: outlier addition {outlier}')
        i += 1

# resetting the X & Y values to initial once (ie, removing outliers)
    X = b * np.sin(phi)
    Y = a * np.cos(phi)
```





Observation

- Used SGDRegressor with `loss='squared_error'`, ie default one.
- When the regulrization term alpha becomes 0.001 and 1, the hyperplane is bending towards the ourliers.
- With low alpha values this is tring to classify outlier points and this can lead to overfitting of the model.
- For `outlier (0,2)`, there is a positive slope for `alpha = 100` and while for rest alpha values the slope is negative
- As `alpha` value increasing there is a reduction in slope for the regression line.
- For `alpha = 100` the slope is small comparing to the rest values and plots and the hyperplane is not tring to classify the outlier points.
  - This also helps us not to overfit.
  - The change in slope will reflect in the prediction phase

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js