CSE 512 Spring 2021 - Machine Learning - Homework 1

Your Name: Irfan Ahmed

Solar ID: 113166464

NetID email address: irfan.ahmed@stonybrook.edu

1) X_1, X_2 are continuous independent R.V $\sim U(0,1)$

$X = \max(X_1, 2X_2)$

i) $E(X) = \int_0^2 x \Pr(X=x)\, dx$

First let us calculate cdf of x.

$F_X(x) = \Pr(X \le x) \quad \forall \; 0 < x < 1.$

$= \Pr(\max\{X_1, 2X_2\} \le x) = \Pr(X_1 \le x \cap 2X_2 \le x)$

$= \Pr(X_1 \le x) \cdot \Pr(2X_2 \le x)$ [$\because$ X_1, X_2 are independent]

$= x \cdot x/2 = x^2/2$

$\forall \; 1 < x < 2$

$F_X(x) = \Pr(X_1 \le x) \cdot \Pr(X_2 \le x/2) = 1 \cdot x/2 = x/2$

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x^2/2 & 0 < x < 1 \\ x/2 & 1 < x < 2 \\ 1 & x > 2 \end{cases}$$

$f_X(x)$ pdf of X $= F'_X(x)$ $= \begin{cases} 0 & x < 0 \\ x & 0 < x < 1 \\ 1/2 & 1 < x < 2 \end{cases}$

$$E(X) = \int_0^2 x P(X=x)\, dx = \int_0^1 x P(X=x)\, dx + \int_1^2 x P(X=x)\, dx$$

$$= \int_0^1 x(x)\, dx + \int_1^2 x(1/2)\, dx$$

$$= \frac{x^3}{3}\Big|_0^1 + \frac{x^2}{4}\Big|_1^2$$

$$= 1/3 + (1 - 1/4) = 1/3 + 3/4 = \frac{13}{12} = 1.0833$$

11) $\text{Var}(X) = E(X^2) - E(X)^2$

$$E(X^2) = \int_0^2 x^2 P(X=x)\,dx = \int_0^1 x^2 P(X=x)\,dx + \int_1^2 x^2 P(X=x)\,dx$$

$$= \int_0^1 x^2 \cdot x\,dx + \int_1^2 x^2 (1/2)\,dx$$

$$= \frac{x^4}{4}\Big|_0^1 + \frac{x^3}{6}\Big|_1^2$$

$$= 1/4 + 8/6 - 1/6 = 1/4 + 7/6 = \frac{6+28}{24} = \frac{34}{24}$$

$$\text{Var}(X) = \frac{34}{24} - \left(\frac{13}{12}\right)^2 = 0.243056$$

III) $\mathrm{Cov}(X_1, X) = E(XX_1) - E(X)E(X_1)$

Calculating cdf of (X, X_1)

$F_{XX_1} = \Pr(X_1 \le x_1, X \le x)$

We know that (Total Probability)

$$\Pr(X \le x) = \Pr(X_1 \le x_1, X \le x) + \Pr(X_1 > x_1, X \le x) \quad — (1)$$

$\Pr(X_1 \le x_1, X \le x) = \Pr(X \le x) - \Pr(X_1 > x_1, X \le x).$

We know $X = \max\{X_1, 2X_2\}$, therefore $X \geq X_1$ always.

$\forall\ 0 < x < 1,\ 0 < x_1 < 1$

$$\Pr(X_1 > x_1, X \le x) = \Pr(x_1 < X_1 \le x, 2X_2 \le x)$$
$$= \cancel{\Pr(}\ (x - x_1)\cdot x/2$$

$$\therefore \Pr(X_1 \le x_1, X \le x) = x^2/2 - (x - x_1)\cdot x/2 = xx_1/2.$$

$\forall\ 1 < x < 2,\ 0 < x_1 < 1$ ($\because$ when $x \in (1,2)$. $X = 2X_2$).

$$\Pr(X_1 \le x_1, X \le x) = \Pr(X_1 \le x_1, 2X_2 \le x) = x_1 \cdot x/2$$

$$F_{XX_1} = \begin{cases} 0 & x < 0,\ x_1 < 0 \\ xx_1/2 & 0 < x < 1,\ 0 < x_1 < 1 \\ xx_1/2 & 1 < x < 2,\ 0 < x_1 < 1 \\ 1 & x > 2,\ x > 1 \end{cases}$$

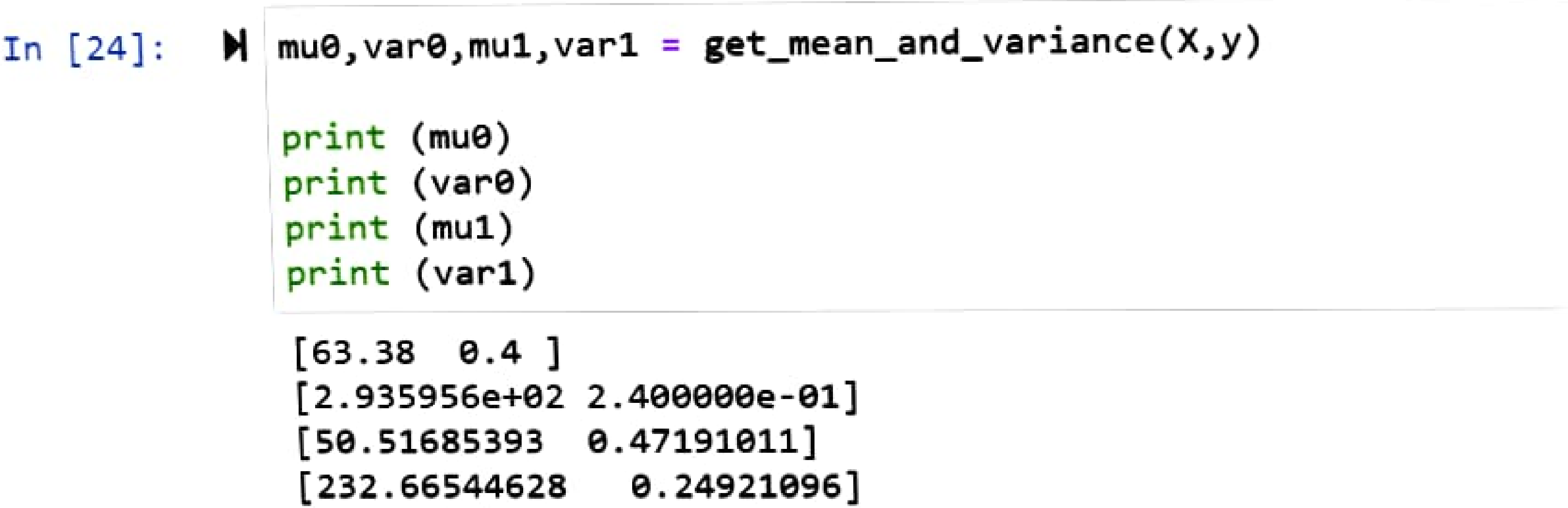
$$f_{XX_1} = \begin{cases} 1/2 & 0 < x < 1,\ 0 < x_1 < 1 \\ 1/2 & 1 < x < 2,\ 0 < x_1 < 1 \end{cases}$$

$$E(XX_1) = \int_0^2 \int_0^1 xx_1 \Pr(X = x \cap X_1 = x_1)\, dx\, dx_1 = \int_0^2 \int_0^1 xx_1 (1/2)\, dx_1\, dx$$

$$E(XX_1) = \frac{1}{2}\int_0^2 x\cdot\left(\frac{x_1^2}{2}\right)\Big|_0^1 = \frac{1}{2}\left(\frac{1}{2}\right)\left[\frac{x^2}{2}\right]_0^2 = \frac{1}{4}\times\frac{1}{2}(4-0) = 1/2$$

$$\mathrm{cov}(XX_1) = E(XX_1) - E(X)E(X_1)$$

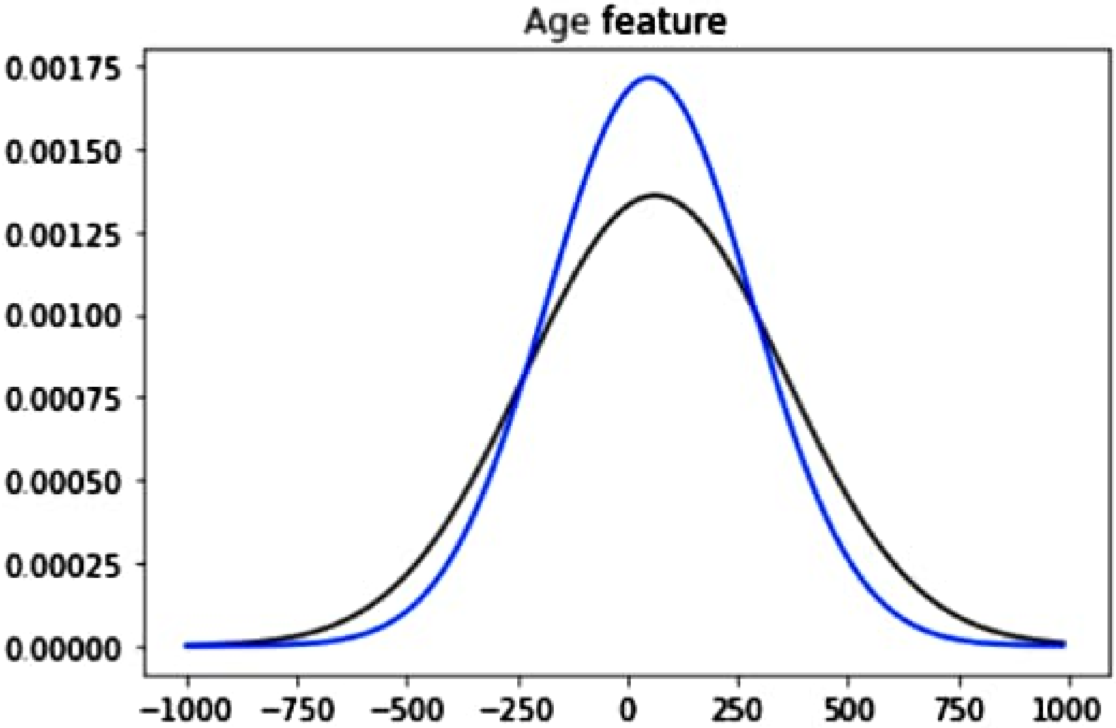
$$= \frac{1}{2} - \left(\frac{13}{12}\right)\left(\frac{1}{2}\right) = \frac{1}{2} - \frac{13}{24} = -1/24$$

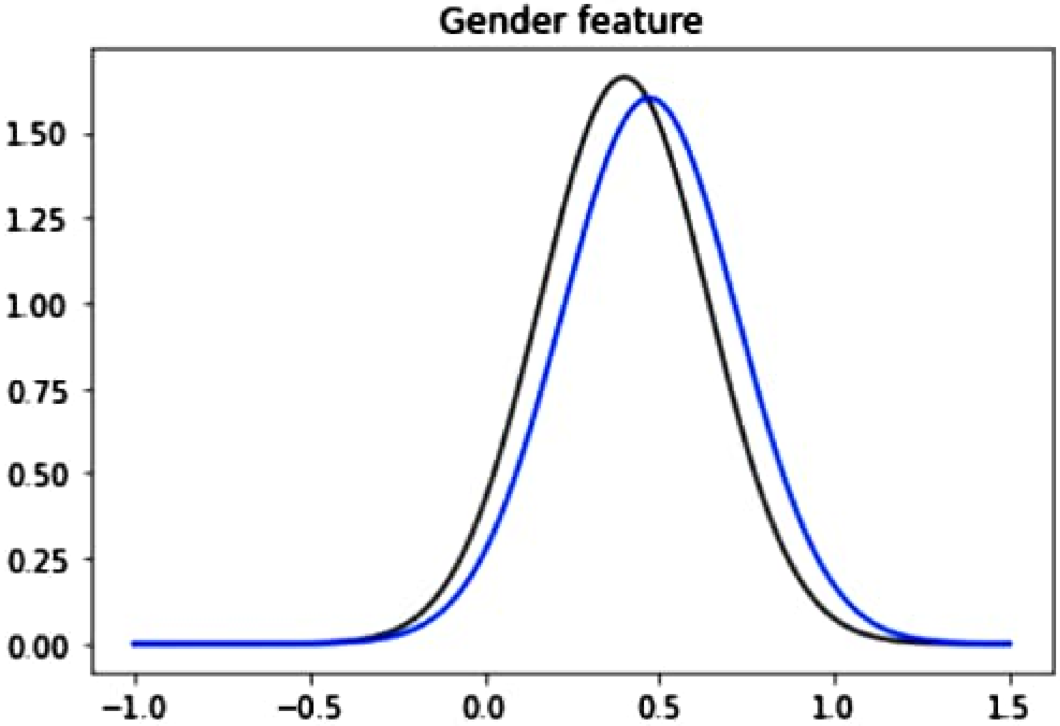


In [24]:

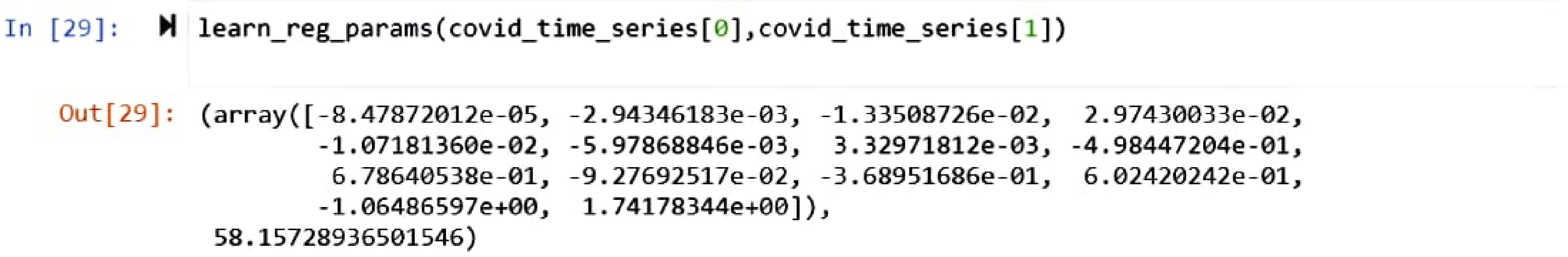
```
mu0,var0,mu1,var1 = get_mean_and_variance(X,y)

print (mu0)
print (var0)
print (mu1)
print (var1)
```

```
[63.38  0.4 ]
[2.935956e+02 2.400000e-01]
[50.51685393  0.47191011]
[232.66544628   0.24921096]
```


Age feature

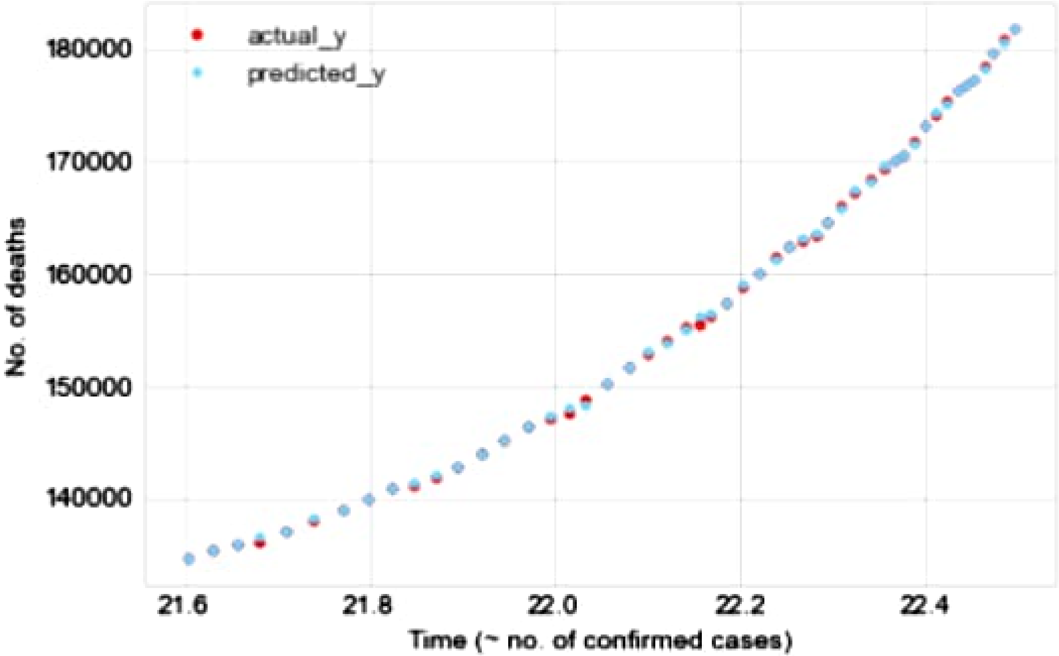

Gender feature

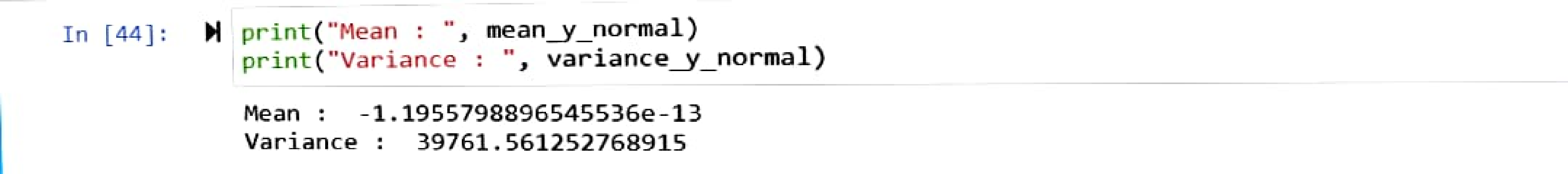
2.2.c :- Approximating gender by Gaussian curve is not a good idea. From the curve we observe we have negative values which is not possible in real life when we have two genders (0 and 1). Also, the decimals do not make any sense. The approximated data gives us physically impossible predictions with non-zero probability. Also, we already know that the population will lie under those two genders giving no additional information from the graph.



```
In [29]:  learn_reg_params(covid_time_series[0],covid_time_series[1])
```

```
Out[29]: (array([-8.47872012e-05, -2.94346183e-03, -1.33508726e-02,  2.97430033e-02,
                -1.07181360e-02, -5.97868846e-03,  3.32971812e-03, -4.98447204e-01,
                 6.78640538e-01, -9.27692517e-02, -3.68951686e-01,  6.02420242e-01,
                -1.06486597e+00,  1.74178344e+00]),
          58.15728936501546)
```

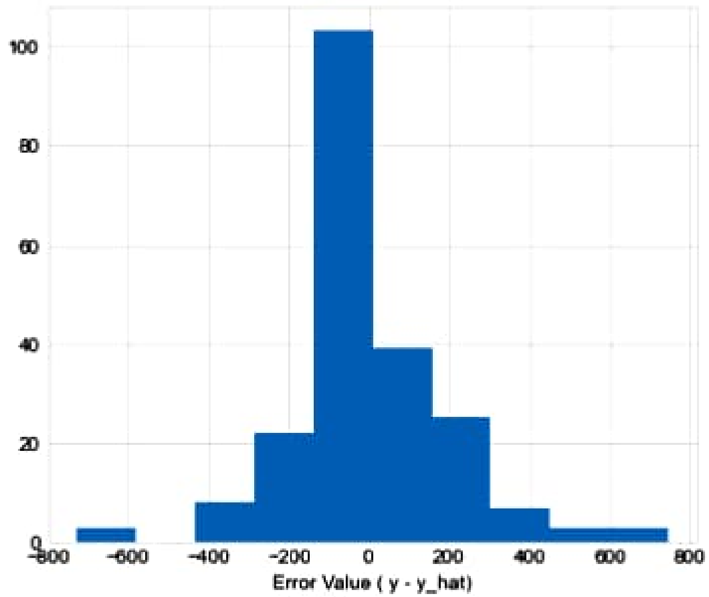

actual_y
predicted_y
No. of deaths
180000
170000
160000
150000
140000
21.6
21.8
22.0
22.2
22.4
Time (~ no. of confirmed cases)

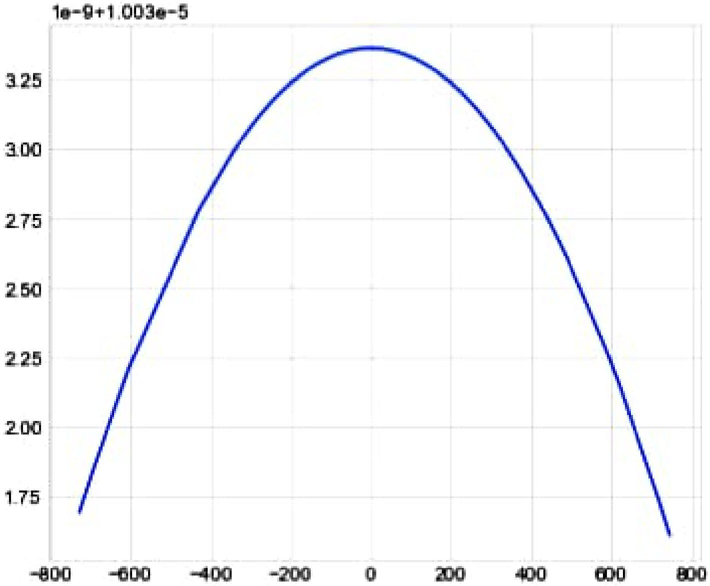


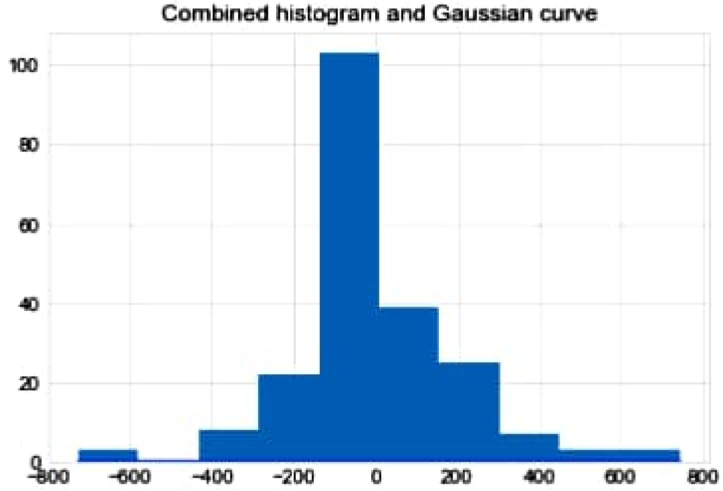
```
print("Mean : ", mean_y_normal)
print("Variance : ", variance_y_normal)
```

```
Mean :  -1.1955798896545536e-13
Variance :  39761.561252768915
```

```
Mean :  -1.1955798896545536e-13
Variance :  39761.561252768915
```






Combined histogram and Gaussian curve

Yes, gaussian is a good approximation for errors. From the histogram, it can be seen that the errors gather around a value(~0) and as per Central Limit Theorem, it looks that the data(i.e. error values) has normal distribution.