Linear Regression

In [3]:
```
!pip install hvplot
```

```
Collecting hvplot
  Downloading hvplot-0.9.2-py2.py3-none-any.whl (1.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 9.7 MB/s eta 0:00:00
Requirement already satisfied: bokeh>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (3.3.4)
Requirement already satisfied: colorcet>=2 in /usr/local/lib/python3.10/dist-packages (from hvplot) (3.1.0)
Requirement already satisfied: holoviews>=1.11.0 in /usr/local/lib/python3.10/dist-packages (from hvplot)
(1.17.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from hvplot) (2.0.3)
Requirement already satisfied: numpy>=1.15 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.25.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from hvplot) (24.0)
Requirement already satisfied: panel>=0.11.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.3.
8)
Requirement already satisfied: param<3.0,>=1.12.0 in /usr/local/lib/python3.10/dist-packages (from hvplot)
(2.1.0)
Requirement already satisfied: Jinja2>=2.9 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hv
plot) (3.1.3)
Requirement already satisfied: contourpy>=1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->h
vplot) (1.2.1)
Requirement already satisfied: pillow>=7.1.0 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->
hvplot) (9.4.0)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->h
vplot) (6.0.1)
Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->h
vplot) (6.3.3)
Requirement already satisfied: xyzservices>=2021.09.1 in /usr/local/lib/python3.10/dist-packages (from bokeh
>=1.0.0->hvplot) (2024.4.0)
Requirement already satisfied: pyviz-comms>=0.7.4 in /usr/local/lib/python3.10/dist-packages (from holoviews
>=1.11.0->hvplot) (3.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from panda
s->hvplot) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot)
(2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplo
t) (2024.1)
Requirement already satisfied: markdown in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvpl
ot) (3.6)
Requirement already satisfied: markdown-it-py in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0
->hvplot) (3.0.0)
Requirement already satisfied: linkify-it-py in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0-
>hvplot) (2.0.3)
Requirement already satisfied: mdit-py-plugins in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.
0->hvplot) (0.4.0)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvpl
```

```
ot) (2.31.0)
Requirement already satisfied: tqdm>=4.48.0 in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->
hvplot) (4.66.2)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplo
t) (6.1.0)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.10/dist-packages (from panel>=0.1
1.0->hvplot) (4.11.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from Jinja2>=2.9-
>bokeh>=1.0.0->hvplot) (2.1.5)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.
8.2->pandas->hvplot) (1.16.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-packages (from bleach->panel>=
0.11.0->hvplot) (0.5.1)
Requirement already satisfied: uc-micro-py in /usr/local/lib/python3.10/dist-packages (from linkify-it-py->p
anel>=0.11.0->hvplot) (1.0.3)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.10/dist-packages (from markdown-it-py->p
anel>=0.11.0->hvplot) (0.1.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from req
uests->panel>=0.11.0->hvplot) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->panel
>=0.11.0->hvplot) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests-
>panel>=0.11.0->hvplot) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests-
>panel>=0.11.0->hvplot) (2024.2.2)
Installing collected packages: hvplot
Successfully installed hvplot-0.9.2
```

In [17]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import hvplot.pandas

from sklearn.model_selection import train_test_split

from sklearn import metrics

from sklearn.linear_model import LinearRegression

%matplotlib inline

df = pd.read_csv('/content/data/life_expectancy_clean.csv')
```

In [4]:
```
df.head()
```

Out[4]:

	Unnamed: 0	country	year	status	life_expectancy	adult_mortality	infant_deaths	alcohol	percentage expenditure	measles	under_five_d
0	0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	1154	
1	1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	492	
2	2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	430	
3	3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	2787	
4	4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	3013	

In [5]:
```
df.shape
```

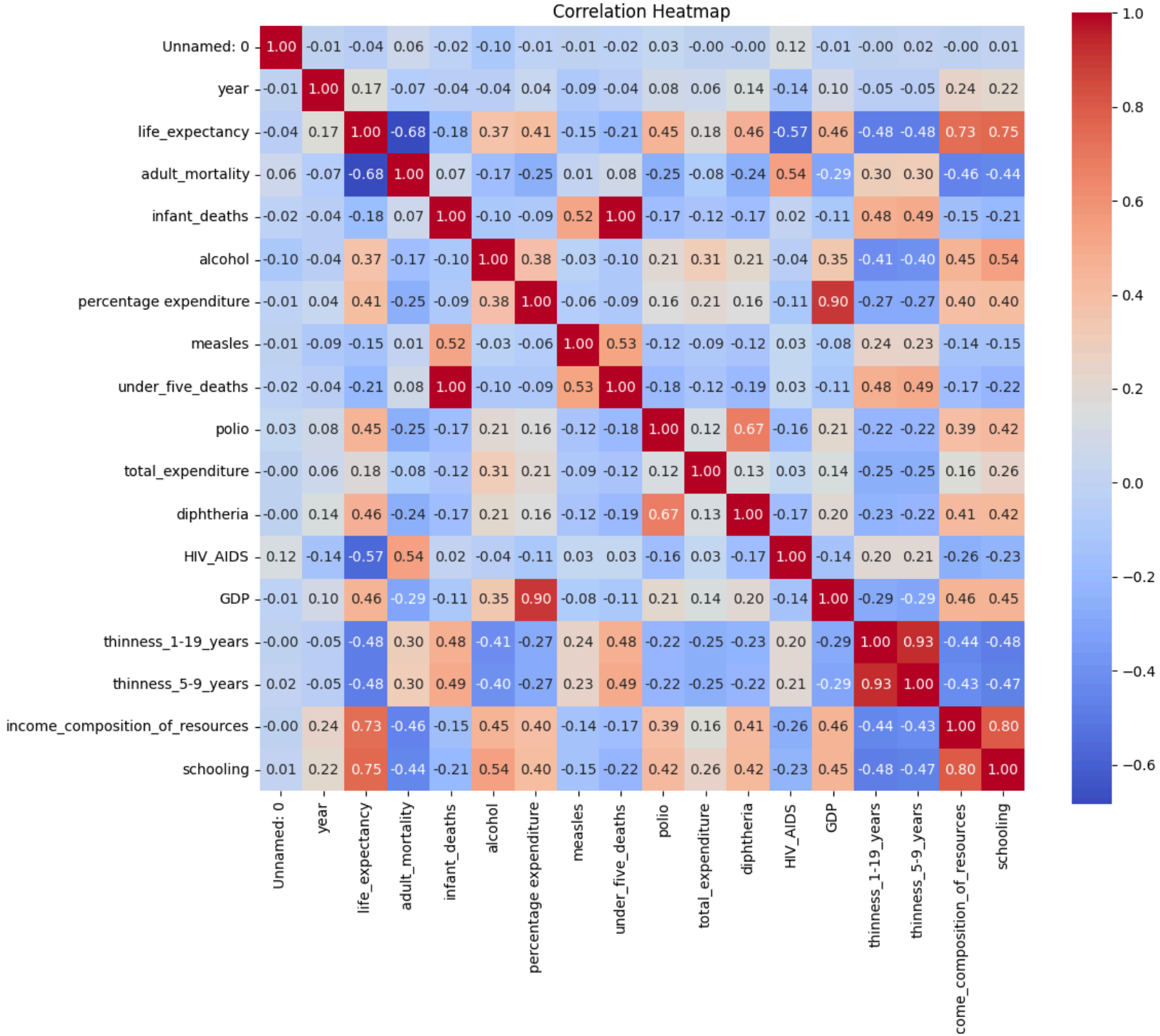
Out[5]:
```
(2480, 20)
```

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2480 entries, 0 to 2479
Data columns (total 20 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Unnamed: 0                       2480 non-null   int64  
 1   country                          2480 non-null   object 
 2   year                             2480 non-null   int64  
 3   status                           2480 non-null   object 
 4    life_expectancy                 2480 non-null   float64
 5    adult_mortality                 2480 non-null   float64
 6   infant_deaths                    2480 non-null   int64  
 7   alcohol                          2480 non-null   float64
 8   percentage expenditure           2480 non-null   float64
 9   measles                          2480 non-null   int64  
 10  under_five_deaths                2480 non-null   int64  
 11  polio                            2480 non-null   float64
 12  total_expenditure                2480 non-null   float64
 13  diphtheria                       2480 non-null   float64
 14  HIV_AIDS                         2480 non-null   float64
 15  GDP                              2480 non-null   float64
 16  thinness_1-19_years              2480 non-null   float64
 17  thinness_5-9_years               2480 non-null   float64
 18  income_composition_of_resources  2480 non-null   float64
 19  schooling                        2480 non-null   float64
dtypes: float64(13), int64(5), object(2)
memory usage: 387.6+ KB
```

In [13]:
```
num_cols = df.select_dtypes(include=['float64', 'int64']).columns
correlation_matrix = df[num_cols].corr()

plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, fmt='.2f', cmap='coolwarm', square=True)
plt.title('Correlation Heatmap')
plt.show()
```


Correlation Heatmap
Unnamed: 0 1.00 -0.01 -0.04 0.06 -0.02 -0.10 -0.01 -0.01 -0.02 0.03 -0.00 -0.00 0.12 -0.01 -0.00 0.02 -0.00 0.01
year -0.01 1.00 0.17 -0.07 -0.04 -0.04 0.04 -0.09 -0.04 0.08 0.06 0.14 -0.14 0.10 -0.05 -0.05 0.24 0.22
life_expectancy -0.04 0.17 1.00 -0.68 -0.18 0.37 0.41 -0.15 -0.21 0.45 0.18 0.46 -0.57 0.46 -0.48 -0.48 0.73 0.75
adult_mortality 0.06 -0.07 -0.68 1.00 0.07 -0.17 -0.25 0.01 0.08 -0.25 -0.08 -0.24 0.54 -0.29 0.30 0.30 -0.46 -0.44
infant_deaths -0.02 -0.04 -0.18 0.07 1.00 -0.10 -0.09 0.52 1.00 -0.17 -0.12 -0.17 0.02 -0.11 0.48 0.49 -0.15 -0.21
alcohol -0.10 -0.04 0.37 -0.17 -0.10 1.00 0.38 -0.03 -0.10 0.21 0.31 0.21 -0.04 0.35 -0.41 -0.40 0.45 0.54
percentage expenditure -0.01 0.04 0.41 -0.25 -0.09 0.38 1.00 -0.06 -0.09 0.16 0.21 0.16 -0.11 0.90 -0.27 -0.27 0.40 0.40
measles -0.01 -0.09 -0.15 0.01 0.52 -0.03 -0.06 1.00 0.53 -0.12 -0.09 -0.12 0.03 -0.08 0.24 0.23 -0.14 -0.15
under_five_deaths -0.02 -0.04 -0.21 0.08 1.00 -0.10 -0.09 0.53 1.00 -0.18 -0.12 -0.19 0.03 -0.11 0.48 0.49 -0.17 -0.22
polio 0.03 0.08 0.45 -0.25 -0.17 0.21 0.16 -0.12 -0.18 1.00 0.12 0.67 -0.16 0.21 -0.22 -0.22 0.39 0.42
total_expenditure -0.00 0.06 0.18 -0.08 -0.12 0.31 0.21 -0.09 -0.12 0.12 1.00 0.13 0.03 0.14 -0.25 -0.25 0.16 0.26
diphtheria -0.00 0.14 0.46 -0.24 -0.17 0.21 0.16 -0.12 -0.19 0.67 0.13 1.00 -0.17 0.20 -0.23 -0.22 0.41 0.42
HIV_AIDS 0.12 -0.14 -0.57 0.54 0.02 -0.04 -0.11 0.03 0.03 -0.16 0.03 -0.17 1.00 -0.14 0.20 0.21 -0.26 -0.23
GDP -0.01 0.10 0.46 -0.29 -0.11 0.35 0.90 -0.08 -0.11 0.21 0.14 0.20 -0.14 1.00 -0.29 -0.29 0.46 0.45
thinness_1-19_years -0.00 -0.05 -0.48 0.30 0.48 -0.41 -0.27 0.24 0.48 -0.22 -0.25 -0.23 0.20 -0.29 1.00 0.93 -0.44 -0.48
thinness_5-9_years 0.02 -0.05 -0.48 0.30 0.49 -0.40 -0.27 0.23 0.49 -0.22 -0.25 -0.22 0.21 -0.29 0.93 1.00 -0.43 -0.47
income_composition_of_resources -0.00 0.24 0.73 -0.46 -0.15 0.45 0.40 -0.14 -0.17 0.39 0.16 0.41 -0.26 0.46 -0.44 -0.43 1.00 0.80
schooling 0.01 0.22 0.75 -0.44 -0.21 0.54 0.40 -0.15 -0.22 0.42 0.26 0.42 -0.23 0.45 -0.48 -0.47 0.80 1.00
Unnamed: 0
year
life_expectancy
adult_mortality
infant_deaths
alcohol
percentage expenditure
measles
under_five_deaths
polio
total_expenditure
diphtheria
HIV_AIDS
GDP
thinness_1-19_years
thinness_5-9_years
come_composition_of_resources
schooling
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

In [25]:
```
df = df.drop(['country', 'status'], axis=1)
```

Training a Linear Regression Model

X and y arrays

In [26]:
```
X=df.drop('infant_deaths', axis=1)
y=df['under_five_deaths']
```

In [27]:
```
print('X=', X.shape, '\ny=', y.shape)
```

```
X= (2480, 17)
y= (2480,)
```

Train Test Split

In [28]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```

In [29]:
```
X_train.shape
```

Out[29]: (1736, 17)

In [30]:
```
X_test.shape
```

Out[30]: (744, 17)

Linear Regression

In [31]:
```
model = LinearRegression()
```

In [32]:
```
model.fit(X_train, y_train)
```

Out[32]:
```
LinearRegression()
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

Model Evaluation

In [33]:
```
model.coef_
```

Out[33]:
```
array([-3.19696146e-15,  1.84054161e-14, -2.06238206e-14, -6.12959009e-16,
       -2.48958210e-14,  4.77076874e-15, -1.59752587e-18,  1.00000000e+00,
       -5.79834139e-16,  1.43346939e-15, -3.00986569e-16, -9.04050635e-16,
       -7.00319223e-16, -6.87831449e-16,  3.42740538e-16,  3.99255731e-15,
        5.47426141e-17])
```

In [35]:
```
pd.DataFrame(model.coef_, X.columns, columns=['Coeficients'])
```

Out[35]:

	Coeficients
Unnamed: 0	-3.196961e-15
year	1.840542e-14
life_expectancy	-2.062382e-14
adult_mortality	-6.129590e-16
alcohol	-2.489582e-14
percentage expenditure	4.770769e-15
measles	-1.597526e-18
under_five_deaths	1.000000e+00
polio	-5.798341e-16
total_expenditure	1.433469e-15
diphtheria	-3.009866e-16
HIV_AIDS	-9.040506e-16
GDP	-7.003192e-16
thinness_1-19_years	-6.878314e-16
thinness_5-9_years	3.427405e-16
income_composition_of_resources	3.992557e-15
schooling	5.474261e-17

Predictions from our Model

In [36]:
```
y_pred = model.predict(X_test)
```

Regression Evaluation Metrics

In [37]:
```
MAE = metrics.mean_absolute_error(y_test, y_pred)
MSE = metrics.mean_squared_error(y_test, y_pred)
RMSE = np.sqrt(MSE)
```

In [38]:
```
MAE
```

Out[38]:
```
3.4727576348680426e-12
```

In [39]:
```
MSE
```

Out[39]:
```
2.92672844164926e-23
```

In [40]:
```
RMSE
```

Out[40]:
```
5.409924622071236e-12
```

In [41]:
```
df['under_five_deaths'].mean()
```

Out[41]:
```
42.884274193548386
```

Residual Histogram

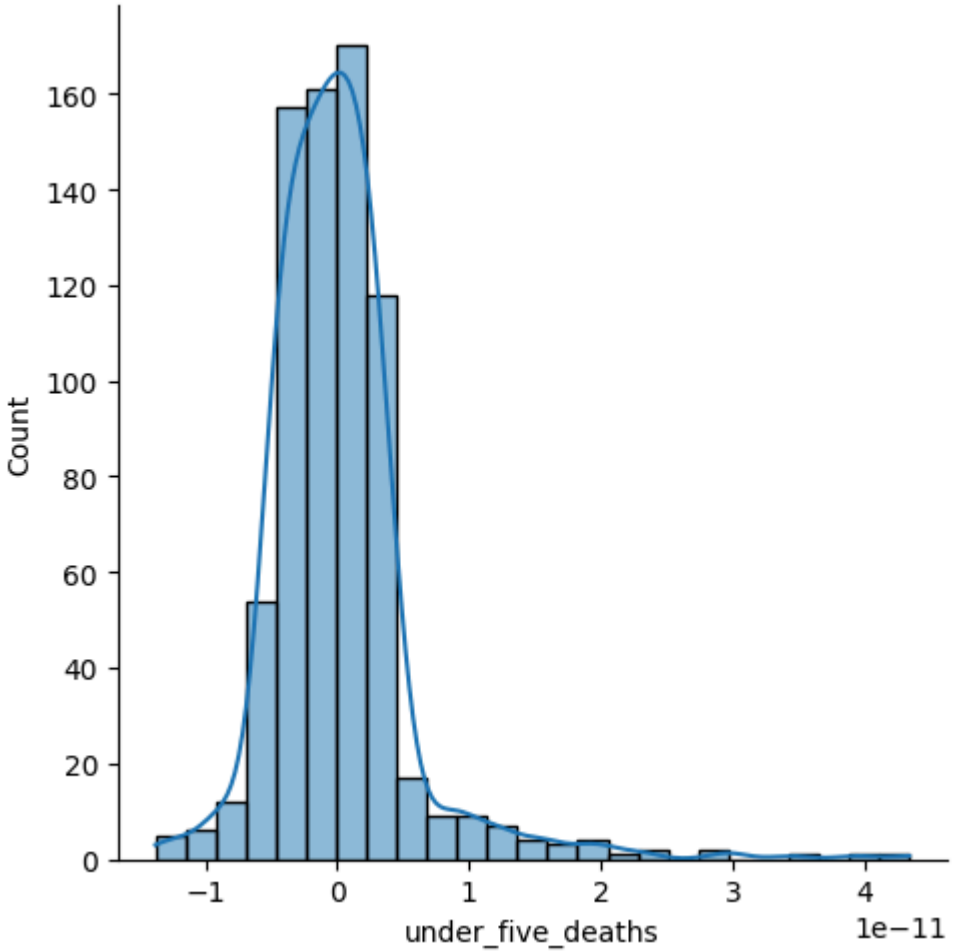
In [42]:
```
test_residual = y_test - y_pred
```

In [45]:
```
pd.DataFrame({'Error Values': (test_residual)}).hvplot.kde()
```

Out[45]:

In [44]:
```
sns.displot(test_residual, bins=25, kde=True)
```

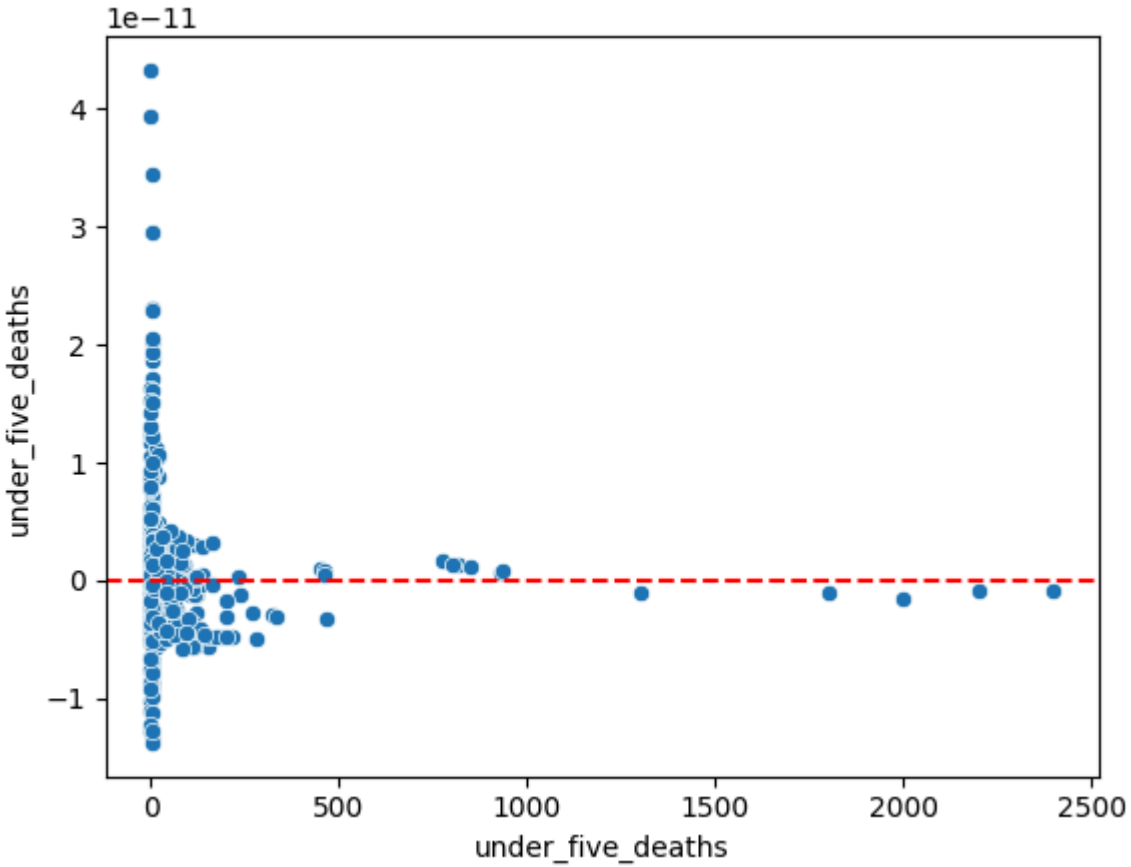
Out[44]:
```
<seaborn.axisgrid.FacetGrid at 0x7c615087bf70>
```



In [46]:
```
sns.scatterplot(x=y_test, y=test_residual)

plt.axhline(y=0, color='r', ls='--')
```

Out[46]: <matplotlib.lines.Line2D at 0x7c6131b012a0>



In [ ]: