De_Guzman_Hands_on_Activity_11_1_Linear_Regression_Analysis Linear Regression

April 24, 2024

1 Hands-on Activity 11.1 Linear Regression Analysis

1.1 Objective(s):

- This activity aims to demonstrate how to apply simple linear regression analysis to solve regression problem

1.2 Intended Learning Outcomes (ILOs):

- Demonstrate how to solve regression problems using simple linear regression
- Use the linear regression model to predict the target value

1.3 Resources:

- Jupyter Notebook

1.4 Files:

- Life Expectancy Data.csv

1.5 Submission Requirements:

- PDF containing initial EDA and Data Wrangling
- PDF showing demonstration of simple linear regression.
- Submit a link to the colab file through the comment section.

1.6 Procedure:

1.6.1 Simple Linear Regression

[100]:
```
!pip install hvplot
```

```
Requirement already satisfied: hvplot in /usr/local/lib/python3.10/dist-packages
(0.9.2)
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->hvplot) (3.1.3)
Requirement already satisfied: contourpy>=1 in /usr/local/lib/python3.10/dist-
packages (from bokeh>=1.0.0->hvplot) (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->hvplot) (6.0.1)
Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.10/dist-
packages (from bokeh>=1.0.0->hvplot) (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 pandas->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->hvplot) (2024.1)
Requirement already satisfied: markdown in /usr/local/lib/python3.10/dist-
packages (from panel>=0.11.0->hvplot) (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->hvplot) (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->hvplot) (6.1.0)
Requirement already satisfied: typing-extensions in
```

```
/usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->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->panel>=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->panel>=0.11.0->hvplot) (0.1.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->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)
```

[101]:
```
import hvplot.pandas
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.linear_model import LinearRegression
```

X and y arrays

[102]:
```
X = life_df_nums.drop('%_expenditure', axis=1)
y= life_df_nums['gdp']
```

[103]:
```
print("X = ", X.shape, "\ny = ", y.shape)
```

```
X =  (2938, 19)
y =  (2938,)
```

Train Test Split

[104]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3,␣
↪random_state = 101)
```

[105]:
```
X_train.shape
```

[105]: (2056, 19)

[106]:
```
X_test.shape
```

[106]:
```
(882, 19)
```

Linear Regression

[107]:
```
model = LinearRegression()
```

[108]:
```
model.fit(X_train, y_train)
```

[108]:
```
LinearRegression()
```

Model Evaluation

[109]:
```
model.coef_
```

[109]:
```
array([ 8.41704544e-13, -1.87296229e-12,  4.34992121e-14,  1.22825409e-13,
        2.81262610e-13, -3.64834892e-14, -1.44751294e-16,  2.29527913e-14,
       -1.05001084e-14,  9.16511064e-15, -1.88127292e-13,  1.57282966e-14,
        1.59484769e-13,  1.00000000e+00, -2.71192083e-18, -2.42147310e-16,
       -1.74052970e-14, -5.45358880e-12,  2.74267897e-13])
```

[110]:
```
pd.DataFrame(model.coef_, X.columns, columns = ['Coefficients'])
```

[110]:

	Coefficients
year	8.417045e-13
life_expectancy	-1.872962e-12
adult_mortality	4.349921e-14
infant_deaths	1.228254e-13
alcohol	2.812626e-13
hepatitis_b	-3.648349e-14
measles	-1.447513e-16
bmi	2.295279e-14
deaths_under_5	-1.050011e-14
polio	9.165111e-15
total_expenditure	-1.881273e-13
diphtheria	1.572830e-14
hiv/aids	1.594848e-13
gdp	1.000000e+00
population	-2.711921e-18
thinness(minors)	-2.421473e-16
thinness(children)	-1.740530e-14
resource_income_composition	-5.453589e-12
schooling	2.742679e-13

Predictions from the Model

[111]:
```
y_pred = model.predict(X_test)
```

Regression Evaluation Metrics

[112]:
```
MAE = metrics.mean_absolute_error(y_test, y_pred)
MSE = metrics.mean_squared_error(y_test, y_pred)
RMSE = np.sqrt(MSE)
```

[113]:
```
MAE
```

[113]: 4.029653104017833e-11

[114]:
```
MSE
```

[114]: 1.4581510424689505e-20

[115]:
```
RMSE
```

[115]: 1.2075392509019948e-10

[116]:
```
life_df_nums['gdp'].mean()
```

[116]: 6393.938563228319

Residual Histogram

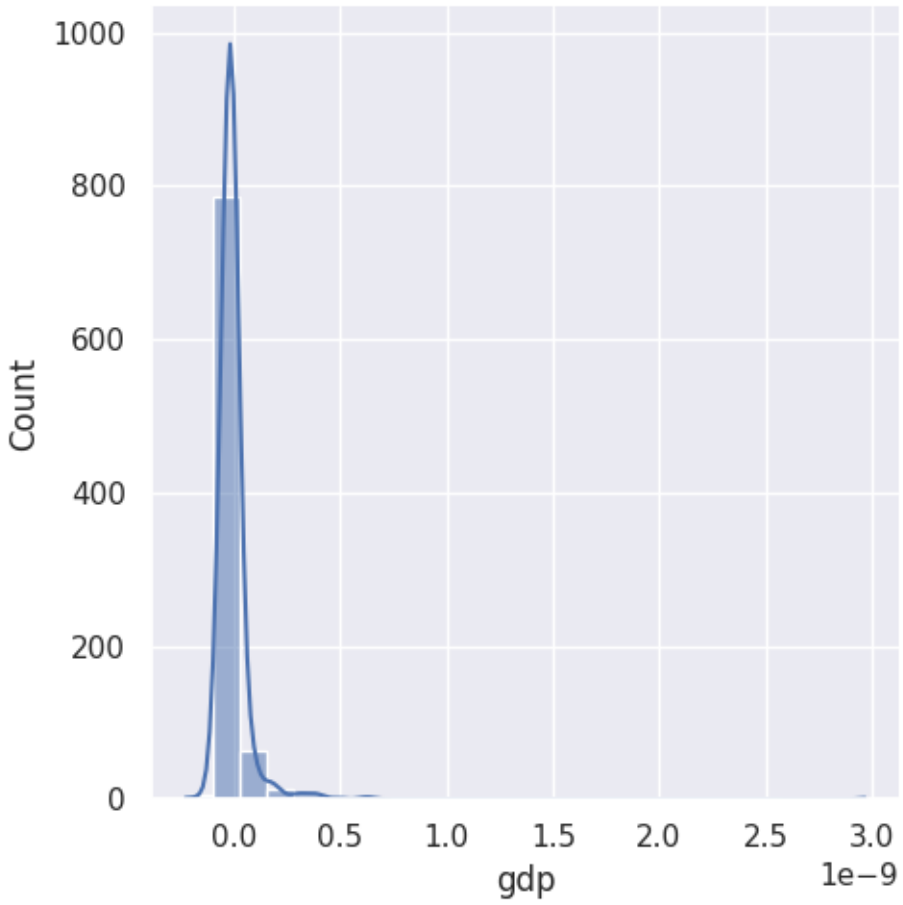
[117]:
```
test_residual = y_test - y_pred
```

[118]:
```
pd.DataFrame({'Error Values': (test_residual)}).hvplot.kde()
```

[118]: :Distribution [Error Values] (Density)

[121]:
```
sns.displot(test_residual, bins=25, kde=True)
```

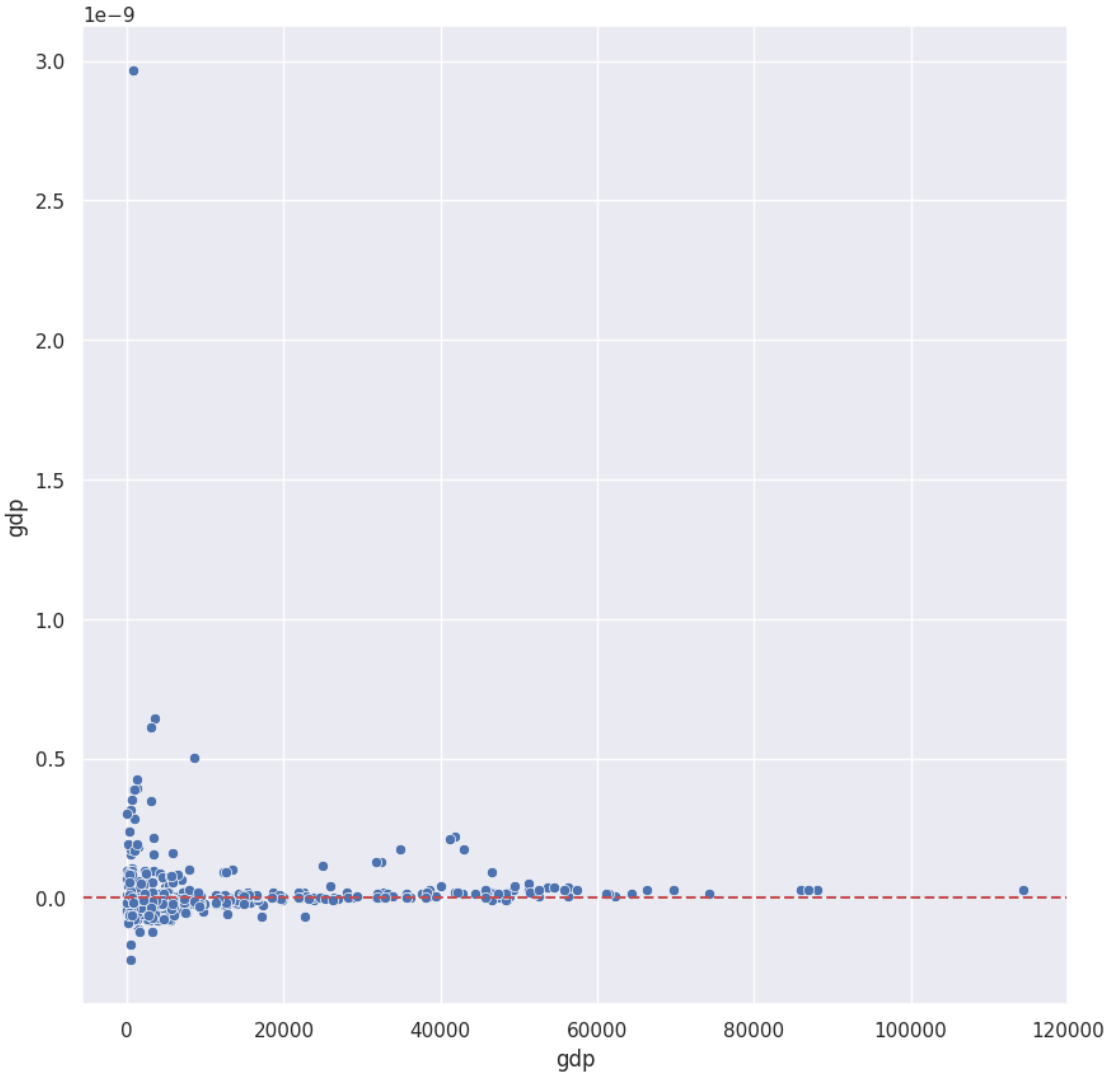
[121]: <seaborn.axisgrid.FacetGrid at 0x78872739b100>



[122]:
```
sns.scatterplot(x=y_test, y=test_residual)

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

[122]: <matplotlib.lines.Line2D at 0x78872164fca0>



1.7 Conclusion

From this activity, I was able to use Linear Regression to compare 2 variables from a dataset and then create a prediction model based on these two. I was able to learn about how we can create a regression model and use it to predict one variable from another. Being able to identify which variables have a really strong correlation with each other allows us to know whether we can construct a consistent and reliable prediction model from linear regression.

[ ]: