Lab 9

Daniel Mehta n01753264

In [12]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso, Ridge
```

In [2]:
```
#Simple Linear
df = pd.DataFrame({
    "Years of Experience": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    "Salary": [30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000]
})

X = df[["Years of Experience"]]
y = df["Salary"]


#Multilinear
df_multi = pd.DataFrame({
    "Years of Experience": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    "Education Level":     [1, 1, 2, 2, 2, 3, 3, 3, 2, 1],
    "Location":            [1, 2, 1, 3, 2, 3, 1, 2, 3, 1],
    "Salary":              [30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000]
})
```

In [3]:
```
df_encoded = pd.get_dummies(df_multi, columns=["Location"], drop_first=True)

X_multi = df_encoded.drop("Salary", axis=1)
y_multi = df_encoded["Salary"]
```

Question 1: Simple Linear Regression

You are given the dataset above, where the independent variable is years_of_experience, and the dependent variable is salary. Implement a simple linear regression model to predict salary based on years of experience. Calculate the regression coefficients and plot the best-fitting line.

In [4]:

```
model = LinearRegression()
model.fit(X, y)

intercept = model.intercept_
slope = model.coef_[0]

print(f"Intercept: {intercept}")
print(f"Slope (Coefficient for Experience): {slope}")

y_pred = model.predict(X)

plt.scatter(X, y, color='blue', label='Actual Data', edgecolor='black')
plt.plot(X, y_pred, color='red', label='Best-Fitting Line')

plt.xlabel("Years of Experience")
plt.ylabel("Salary")
plt.title("Experience vs Salary")
plt.legend()
plt.grid(True)

plt.show()
```

```
Intercept: 24999.99999999999
Slope (Coefficient for Experience): 5000.000000000002
```

Question 2: Interpretation of the Slope and Intercept

Using the simple regression model created in Question 1, explain the meaning of the slope and intercept of the regression line. What do they represent in the context of this dataset?

Intrepretation

- **Intercept (25,000):** This is the predicted salary for an employee with **0 years of experience**. It suggests a starting salary of **$25,000** even with no experience.
- **Slope (5,000):** This means that for **each additional year of experience**, the salary **increases by $5,000**.

For example, if someone had 4 years of experience, there predicted salary would be:

25,000 + (4 * 5,000) = 45,000

Question 3: Multiple Linear Regression

The dataset contains years_of_experience, education_level, and location as independent variables and salary as the dependent variable. Implement a multiple regression model and explain the relationship between these independent variables and salary

In [5]:
```
model_multi = LinearRegression()
model_multi.fit(X_multi, y_multi)

for col, coef in zip(X_multi.columns, model_multi.coef_):
    print(f"{col}: {coef:.2f}")
print(f"Intercept: {model_multi.intercept_:.2f}")
```

```
Years of Experience: 5000.00
Education Level: 0.00
Location_2: 0.00
Location_3: 0.00
Intercept: 25000.00
```

Question 4: Error Estimation in Linear Regression

Using the regression model from Question 1, calculate the Mean Squared Error (MSE) and the RMSE to evaluate the model's performance.

In [6]:
```
mse = mean_squared_error(y, y_pred)
rmse = np.sqrt(mse)
print(f"MSE: {mse}")
print(f"RMSE: {rmse}")
```

```
MSE: 1.588186776101813e-23
RMSE: 3.9852061127397324e-12
```

Question 5: Best Fit Line Calculation

Given the dataset for simple linear regression, calculate the best-fit line using the formula for simple linear regression.

In [7]:
```
x_mean = X.mean()
y_mean = y.mean()

slope_np, intercept_np = np.polyfit(X.values.flatten(), y, 1)

print(f"Slope: {slope_np}")
print(f"Intercept: {intercept_np}")
```

```
Slope: 5000.0
Intercept: 24999.999999999967
```

Question 6: Predicting Values

Using the best-fit line obtained in Question 5 to predict the salary for an employee with 6 years of experience.

In [9]:
```
salary_6_years = intercept_np + slope_np * 6
print(f"Predicted salary for 6 years of experience: ${salary_6_years:.2f}")
```

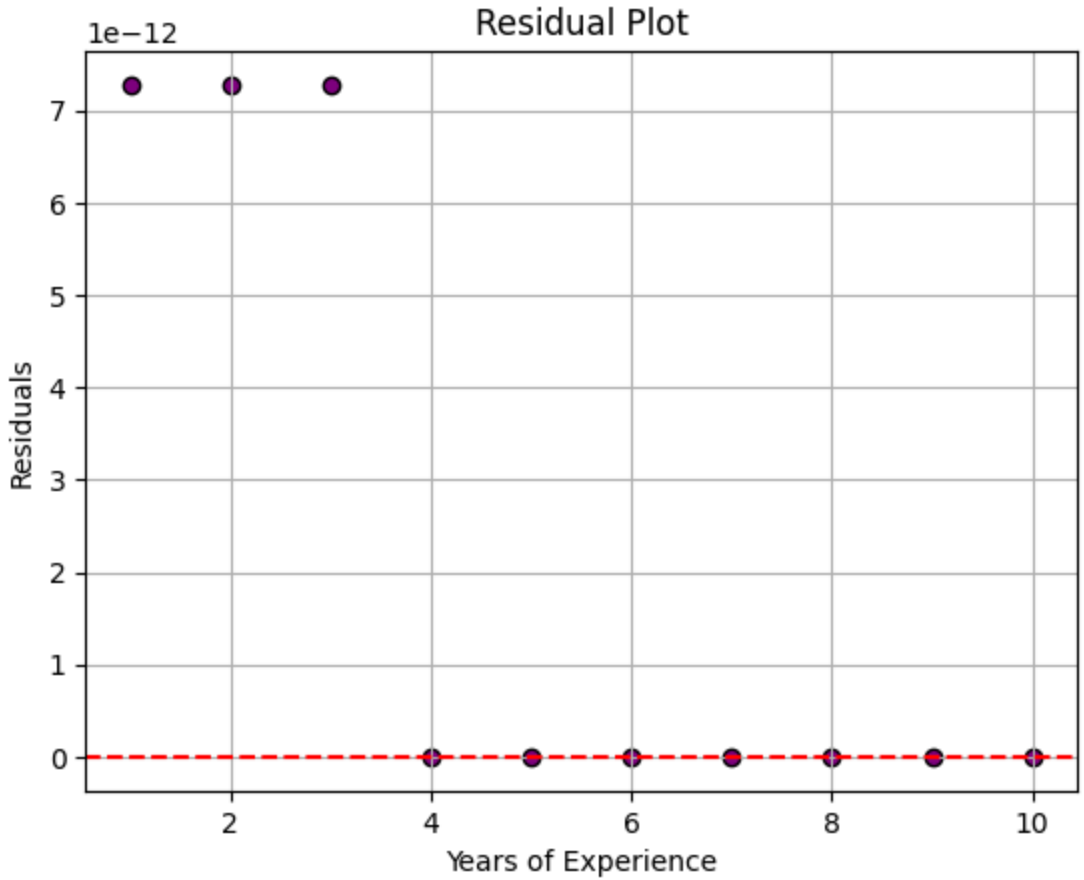
```
Predicted salary for 6 years of experience: $55000.00
```

Question 7: Residuals Analysis

For the simple regression model in Question 1, calculate the residuals (the difference between actual and predicted values). Plot the residuals to check if the assumptions of linear regression are met.

In [10]:
```
residuals = y - y_pred


plt.scatter(X, residuals, color='purple', edgecolor='black')
plt.axhline(y=0, color='red', linestyle='--')
plt.title("Residual Plot")
plt.xlabel("Years of Experience")
plt.ylabel("Residuals")
plt.grid(True)
plt.show()
```



Question 8: Correlation and Regression Coefficients

Using the dataset for multiple linear regression, calculate the correlation coefficient between each independent variable and the dependent variable (salary).

In [11]:
```
correlations = df_multi[["Years of Experience", "Education Level", "Location", "Salary"]].corr()
print(correlations["Salary"].drop("Salary"))
```

```
Years of Experience    1.000000
Education Level        0.359573
Location               0.146695
Name: Salary, dtype: float64
```

Question 9: Regularization in Multiple Regression

In a multiple regression scenario with several predictors, apply Lasso (L1 regularization) and Ridge (L2 regularization) to prevent overfitting. Compare the coefficients obtained using these methods.

In [20]:
```
lasso = Lasso(alpha=0.1)
lasso.fit(X_multi, y_multi)

ridge = Ridge(alpha=0.1)
ridge.fit(X_multi, y_multi)

print("Lasso Coefficients:")
for col, coef in zip(X_multi.columns, lasso.coef_):
    print(f"{col}: {coef:.2f}")
print("-"*30)
print("Ridge Coefficients:")
for col, coef in zip(X_multi.columns, ridge.coef_):
    print(f"{col}: {coef:.2f}")
```

```
Lasso Coefficients:
Years of Experience: 4999.99
Education Level: 0.00
Location_2: -0.00
Location_3: 0.00
------------------------------
Ridge Coefficients:
Years of Experience: 4992.95
Education Level: 8.85
Location_2: -3.81
Location_3: 2.43
```

In [ ]: