A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable. Independent variables are also called explanatory or predictor variables.

Example

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
df=pd.read_csv(r"C:\Users\USER\Downloads\archive\weight-height.csv")
```

In [3]:
```
df.head()
```

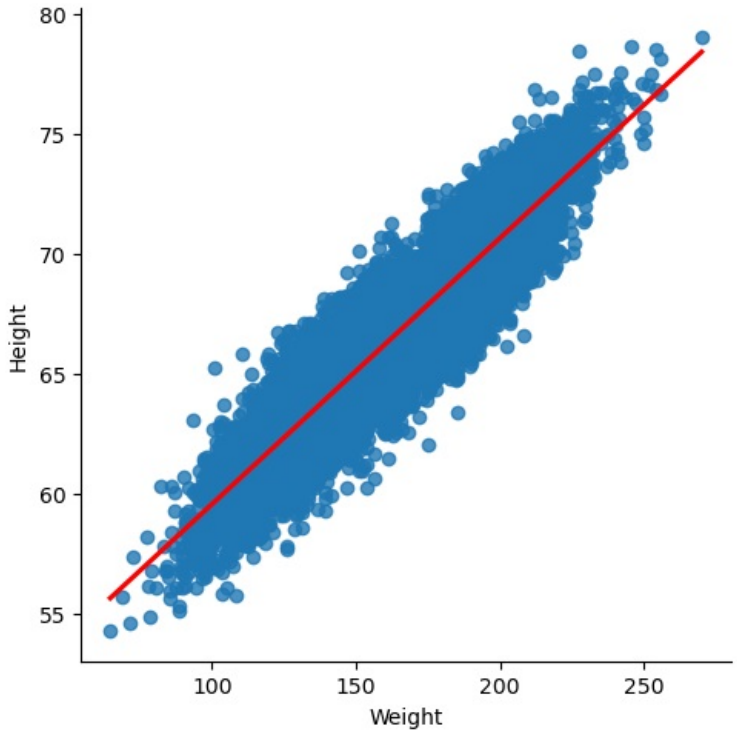
Out[3]:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801

In [10]:
```
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

In [11]:
```
sns.lmplot(x='Weight', y='Height', data=df, line_kws={'color': 'red'})
```

Out[11]:
```
<seaborn.axisgrid.FacetGrid at 0x18ed3e3fad0>
```



In [12]:
```
from sklearn.linear_model import LinearRegression
```

In [15]:
```
X=df[['Weight']]
y=df['Height']
```

In [16]:
```
# Fit the linear regression model
model = LinearRegression()
model.fit(X, y)
```

Out[16]:
```
LinearRegression
LinearRegression()
```

In [17]:
```
# Make predictions
df['predicted_height'] = model.predict(X)
```

In [19]:
```
# Calculate residuals
df['residuals'] = df['Height'] - df['predicted_height']
```

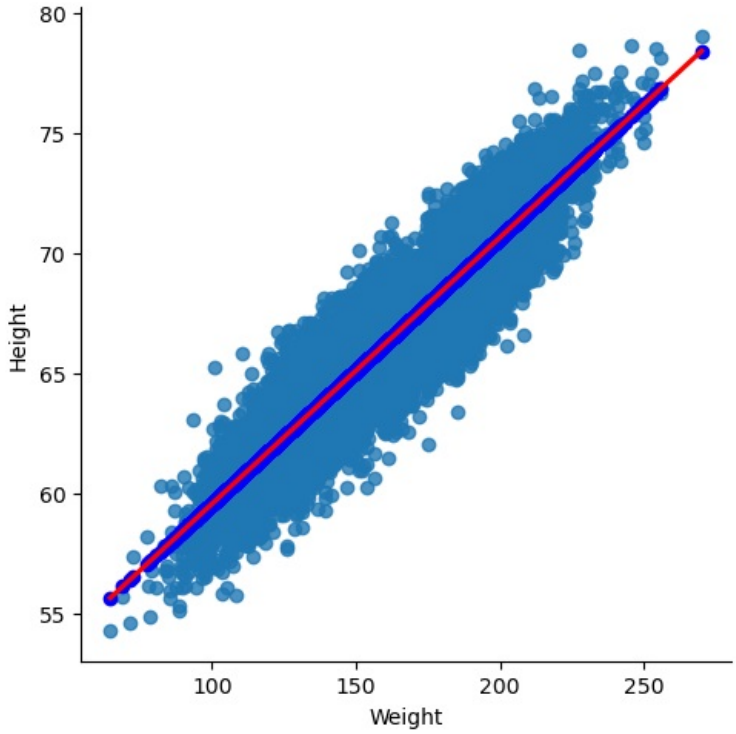
In [20]:
```
df.head()
```

Out[20]:

	Gender	Height	Weight	predicted_height	residuals
0	Male	73.847017	241.893563	75.282804	-1.435787
1	Male	68.781904	162.310473	66.463980	2.317924
2	Male	74.110105	212.740856	72.052311	2.057794
3	Male	71.730978	220.042470	72.861424	-1.130445
4	Male	69.881796	206.349801	71.344101	-1.462305

In [23]:
```
sns.lmplot(x='Weight', y='Height', data=df, line_kws={'color': 'red'})
plt.scatter(df['Weight'], df['predicted_height'], color='blue')
```

Out[23]:
```
<matplotlib.collections.PathCollection at 0x18edba5cd90>
```



So, here predicted_height is best fit line.

The best fit line (or line of best fit) is a straight line that best represents the data on a scatter plot. This line minimizes the sum of the squared differences (residuals) between the observed values and the values predicted by the line