linear regression math

April 26, 2026

[34]:
```
import kagglehub

# Download latest version
path = kagglehub.dataset_download("devbatrax/linear-regression-dataset")

print("Path to dataset files:", path)
```

```
Path to dataset files:
C:\Users\aman9\.cache\kagglehub\datasets\devbatrax\linear-regression-
dataset\versions\1
```

[35]:
```
csv_path = path+'\\linear regression dataset.csv'
```

[36]:
```
import pandas as pd
import numpy as np
```

[37]:
```
df = pd.read_csv(csv_path)
```

[38]:
```
# clearning out nan values
df = df.dropna()
```

[39]:
```
df.head()
```

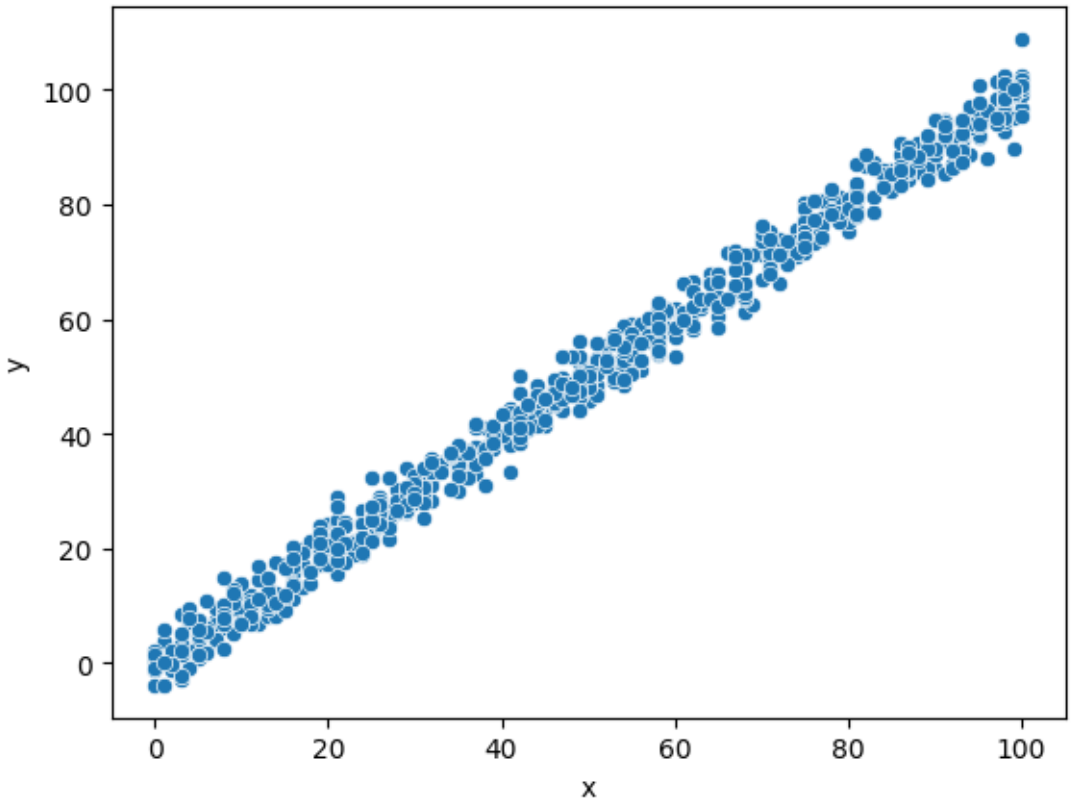
[39]:

	x	y
0	24.0	21.549452
1	50.0	47.464463
2	15.0	17.218656
3	38.0	36.586398
4	87.0	87.288984

[40]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

[41]:
```
sns.scatterplot(x="x", y="y", data=df)
```

[41]:
```
<Axes: xlabel='x', ylabel='y'>
```



[42]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    df['x'], df['y'],
    test_size=0.2,
)
```

[43]:
```
x_mean = X_train.mean()
y_mean = y_train.mean()
```

[44]:
```
covar = 0
for x, y in np.nditer([X_train, y_train]):
    covar += (x-x_mean)*(y-y_mean)
```

[45]:
```
variance_x = 0
for x in X_train:
    variance_x += (x-x_mean)**2
```

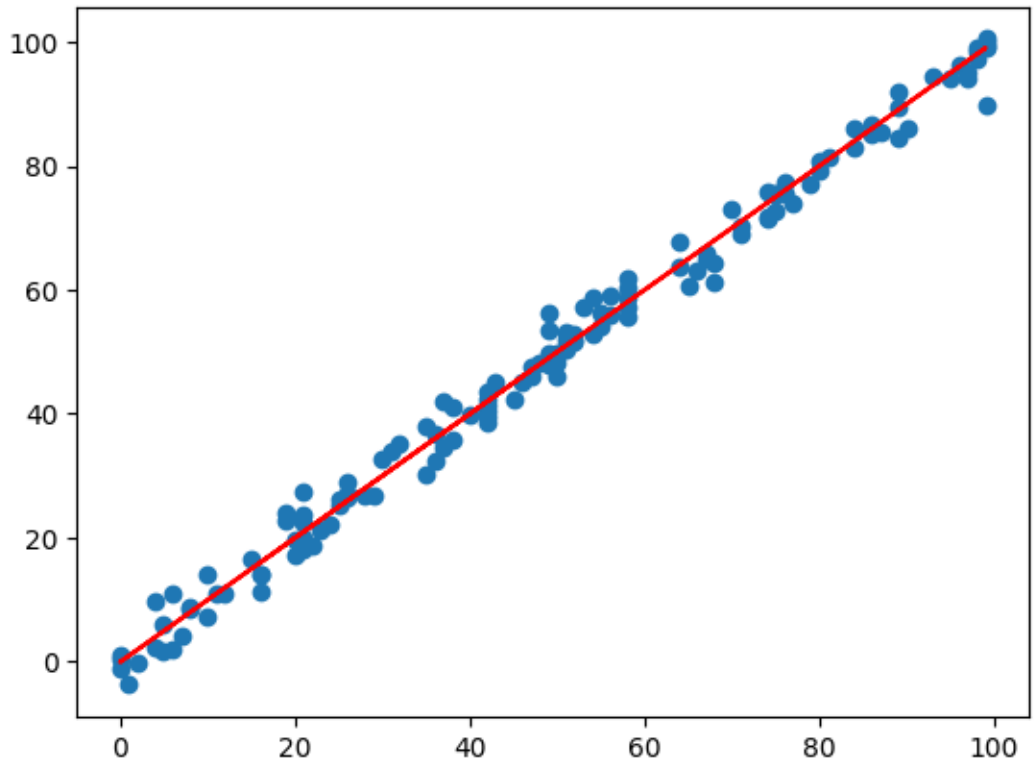
[46]:
```
slope = covar / variance_x
```

[47]:
```
intercept = y_mean - slope*x_mean
```

[48]:
```
y_pred = []
for x in X_test:
    y_pred.append(slope*x + intercept)
y_pred = np.array(y_pred)
```

[51]:
```
plt.scatter(X_test, y_test)
plt.plot(X_test, y_pred, color='red')
```

[51]: [<matplotlib.lines.Line2D at 0x1b84b37ef90>]



[ ]: