Multiple Linear Regression Models

(Multiple) Linear Regression

- In practise there is normally more than one independent variable
- The plane of best fit is given by
-
- $y = a_1 * x_1 + a_2 * x_2 \ldots + a_n * x_n + b$
-
- Again the parameters are obtained by minimising the sum of the squares of the errors.

Multiple Linear regression

$$y = \sum_k a_k x_k + b$$

- y is the dependent variable
- x_k are independent variables
- a_k, b are parameters
- For Example
- StackLoss = a_1*AirFlow + a_2*WaterTemp + a_3*AcidTemp + b

Building the model

```
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

stacklossData = pd.read_csv("data/stackloss.csv")
print(stacklossData)

X = stacklossData[['AirFlow', 'WaterTemp', 'AcidConc']]
y = stacklossData['StackLoss']
X_train, X_test, y_train, y_test =
                                    train_test_split(X, y, random_state=1)

model = LinearRegression()
model.fit(X_train, y_train)

print('intercept:', model.intercept_)
print('slope:', model.coef_)
```

Building the model

- model.fit() calculates the intercept (a) and the slopes (b_1, b_2 etc)
- model.intercept_ and model.coef_ output the following values

intercept: -39.91967442012403
slope: [0.7156402 1.29528612 -0.15212252]

Calculate RMSE

```
yhat = model.predict(X_test)
print(mean_squared_error(y_test, yhat, squared=False))
```

- Finds the predictions for all the test data X_test.
- Finds the RMSE between y_test and yhat.
- Gives an indication of errors.

Making a Prediction

```
# creating a DF from a list of lists, in this case a list of one list
newData = [[72, 20, 85]]
newDF = pd.DataFrame(newData,
                                columns = ['AirFlow', 'WaterTemp', 'AcidConc'])
# print(newDF)

y_hat = model.predict(newDF)
print('y_hat', y_hat)



# output
# y_hat [24.58172837]
```

Interpretation of Coefficients

- (Intercept) Air.Flow Water.Temp Acid.Conc.
- -39.9196744 0.7156402 1.2952861 -0.1521225
-
- If the Air.Flow increases by 1 unit (and other features remain the same) then stack.loss will increase by 0.71 units.
- If the Water.Temp increases by 1 unit (and other features remain the same) then stack.loss will increase by 1.29 units.
- If the Acid.Conc increases by 1 unit (and other features remain the same) then stack.loss will decrease by 0.15 units.

Interpretation of Coefficients

- Note that the size of these coefficients will vary depending on the scale of measurements used.
- For example, if a litre scale is changed to ml, then the corresponding coefficient will increase by a factor of 1000.
- So it is important not to interpret these coefficients as a measure of correlation for example.