Chapter 23

Modeling: Multiple Linear Regression

Multiple regression extends the concept of simple linear regression to analyze the relationship between multiple independent variables and a single dependent variable. While simple linear regression deals with one predictor, multiple regression involves two or more predictors.

In multiple regression, we aim to build a model that predicts the value of a dependent variable based on the values of multiple independent variables.

23.1 Extending the Equation to a Multiple Regression

When there are two x variables, the structure is

$$y = f(x_1, x_2) = m_1 x_1 + m_2 x_2 + b$$

The California housing has eight x variables; therefore, the structure is

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) = m_1 x_1 + m_2 x_2 + m_3 x_3 + m_4 x_4 + m_5 x_5 + m_6 x_6 + m_7 x_7 + m_8 x_8 + b$$

$$y = MedHouseVal$$

$$x_1 = MedInc$$

$$x_2 = HouseAge$$

$$x_3 = AveRooms$$

$$x_4 = AveBedrms$$

$$x_5 = Population$$

$$x_6 = AveOccup$$

$$x_7 = Latitude$$

$$x_8 = Longitude$$

The m_1 through m_8 values are the impact the variable has on the y target value. When all variables are 0, the value of the y variable is the b value.

When there are two or more variables, this function is called a multiple linear regression.

23.2 Implementing the Regression in Python

Step 1: Define the X variable

The idea here is to have the X variable be everything except the target.

```
[96] X = california_df.drop(['MedHouseVal'], axis = 1)
     y = california_df['MedHouseVal']
```


Step 2: Split up the data

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
```


The test size was once again kept at 20%.

Step 3: Building and fitting the model

```
[100] model = LinearRegression()

[101] model.fit(X_train, y_train)

      ▾ LinearRegression
      LinearRegression()
```


Note: When building a multiple regression, there is no need to reshape the values. In this case, as it is a multiple regression, we are not reshaping the values and can build the model directly.

Step 4: Print the model intercept and coefficients

```
model.intercept_ #b value

-37.02327770606409

[103] model.coef_ #slope or m value

      array([ 4.48674910e-01,  9.72425752e-03, -1.23323343e-01,  7.83144907e-01,
             -2.02962058e-06, -3.52631849e-03, -4.19792487e-01, -4.33708065e-01])
```


The equation of the line can be written as follows:

- Replace the m values
  - $y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) = 0.45 * x_1 + 0.0097 * x_2 + -0.123 * x_3 + 0.783 * x_4 + -0.000002 * x_5 + -0.0035 * x_6 + -0.4198 * x_7 + -0.434 * x_8 + b$

- Replace the b value
  - $y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) = 0.45 * x_1 + 0.0097 * x_2 + -0.123 * x_3 + 0.783 * x_4 + -0.000002 * x_5 + -0.0035 * x_6 + -0.4198 * x_7 + -0.434 * x_8 + -37.023$
- Replace the x values
  - $x_1 = MedInc$
  - $x_2 = HouseAge$
  - $x_3 = AveRooms$
  - $x_4 = AveBedrms$
  - $x_5 = Population$
  - $x_6 = AveOccup$
  - $x_7 = Latitude$
  - $x_8 = Longitude$

$$y = 0.45 * MedInc + 0.0097 * HouseAge + -0.123 * AveRooms + 0.783 * AveBedrms + -0.000002 * Population + -0.0035 * AveOccup + -0.4198 * Latitude + -0.434 * Longitude + -37.023$$

- Replace the y value
  - $y = MedHouseVal$

$$MedHouseVal = 0.45 * MedInc + 0.0097 * HouseAge + -0.123 * AveRooms + 0.783 * AveBedrms + -0.000002 * Population + -0.0035 * AveOccup + -0.4198 * Latitude + -0.434 * Longitude + -37.023$$

Step 5: Predict the model

```
[104] y_pred = model.predict(X_test)
```

Step 6: Print the R2 and mean squared error (MSE) values

Q1

```
mse = mean_squared_error(y_pred, y_test)
print(mse)
r2 = r2_score(y_test, y_pred)
print(r2)

0.5558915986952444
0.5757877060324508
```


In this case, this is an average model with low r^2 and high MSE relative to the Median Value of the house.

23.3 Conclusion

In conclusion, we extended our analysis to multiple regression, which allows for the consideration of multiple independent variables. Multiple regression enabled us to examine how several predictors collectively influence the dependent variable. By incorporating multiple predictors, we can create more sophisticated models that better capture the complexities of real-world relationships.

Query
Q1 : page 145 *AU: We have update the number sequence as Step 6. Kindly confirm it is ok?*