Multiple Linear regression

Your code is a good foundation for building a multiple regression model to predict the profit of startups based on various parameters. Here are additional checks and steps you can take to validate and improve your model:

Additional Checks

1. **Check for Multicollinearity**:
   - **What it is**: Multicollinearity occurs when independent variables are highly correlated with each other, which can affect the stability and interpretation of the coefficients.
   - **How to check**: Calculate the Variance Inflation Factor (VIF) for each independent variable.
2. **Check Residuals for Homoscedasticity**:
   - **What it is**: Ensures that the residuals have constant variance.
   - **How to check**: Plot the residuals vs. fitted values.
3. **Check Residuals for Normality**:
   - **What it is**: Ensures that the residuals are normally distributed.
   - **How to check**: Use a Q-Q plot
4. **Check for Independence of Errors**:
   - **What it is**: Ensures that the residuals are independent of each other.
   - **How to check**: Use the Durbin-Watson test.

Model Validation

1. **Cross-Validation**:
   - **What it is**: Helps ensure that the model generalizes well to unseen data.
   - **How to perform**: Use k-fold cross-validation.
2. **Performance Metrics**:
   - **Mean Squared Error (MSE)**: Lower MSE indicates better model performance.
   - **Mean Absolute Error (MAE)**: Lower MAE indicates better model performance.
   - **R-squared (R^2)**: Values closer to 1 indicate that a higher proportion of the variance in the dependent variable is explained by the model.
3. **Learning Curves**:
   - **What it is**: Shows how the model's performance changes with the size of the training data.
   - **How to plot**:

If Checks Fail

If any of these checks fail, consider the following options:

1. **Transform Variables**: Apply transformations (log, square root) to the dependent or independent variables to meet assumptions.
2. **Remove Multicollinearity**: Remove or combine highly correlated variables.
3. **Use Regularization**: Apply techniques like Ridge or Lasso regression to handle multicollinearity and overfitting.
4. **Try Different Models**: Use other models like decision trees, random forests, or neural networks if the relationship is complex.
5. **Add More Features**: Include additional relevant variables to capture more information.
6. **Polynomial Regression**: If the relationship is non-linear, consider polynomial regression.

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

Multiple correlation coefficient R
Measures the relationship/correlation between the dependent variable and the independent variables

Coefficient of determination R^2
Indicates how much of the variance of the dependent variable can be explained by the independent variables

Model summary

Copy

R	R^2	Adjusted R^2	Standard error of the estimate
0.7	0.49	0.299	551.608

Adjusted R^2
R^2 overestimates the coefficient of determination just when many independent variables are used

Standard estimation error
Indicates by how much the model overestimates the dependent variable on average

- F-test to test the null hypothesis, whether the **variance explanation R^2** in the population is zero.

ANOVA

Copy

Model	df	F	p-value
Regression	3	2.561	0.104

- The test is often not of great interest.
- The test is equivalent to asserting that all true slope coefficients in the population are zero.

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

Coefficients

Copy

Model	Unstandardized Coefficients B	Standardized Coefficients Beta	Standard error	t	p-value
(Constant)	1,516.397		1,491.95	1.016	0.339
Male	235.615	0.187	568.819	0.414	0.69
Age	34.726	0.545	27.486	1.263	0.242
Weight	-9.005	-0.178	13.949	-0.646	0.537

Age has the greatest influence on salary

Values less than 0.05 are considered significant

Salary = 1516.397 - 235.615 · **Male** + 34.726 · **Age** – 9.005 · **Weight**

Multicollinearity diagnosis

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$\hat{x}_1 = b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$\hat{x}_2 = b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a$$

$$\hat{x}_k = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + a$$

Tolerance

$$T = 1 - R^2$$

Coefficient of determination

Attention:

$$T < 0.1$$

VIF (Variance Inflation Factor)

$$VIF = \frac{1}{1 - R^2}$$

Coefficient of determination

Attention:

$$VIF > 10$$

Categorical variables

Categorical variables with two characteristics can be used as independent variables (predictors).

Dichotomous, e.g. gender with the characteristics **male** and **female**

0 = female
1 = male

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

variable gender

$$\hat{y} = b_1 \cdot 0 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

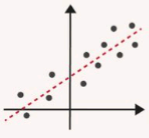
$$\hat{y} = b_1 \cdot 1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

The slope is the difference

Linear regression VS Logistic regression

Linear regression

- Econometric modeling
- Marketing mix model
- Customer lifetime value



Continuous > Continuous

$$y = a_0 + \sum_{i=1}^{N} a_i x_i$$

lm(y~x1 + x2, data)

1 unit increase in x increases y by a

Logistic regression

- Customer choice model
- Click-through rate
- Conversion rate
- Credit scoring

Continuous > True/False

$$y = \frac{1}{1 + e^{-z}}$$

$$z = a_0 + \sum_{i=1}^{N} a_i x_i$$

glm(y~x1 + x2, data), family = binomial())

1 unit increase in x increases log odds by a