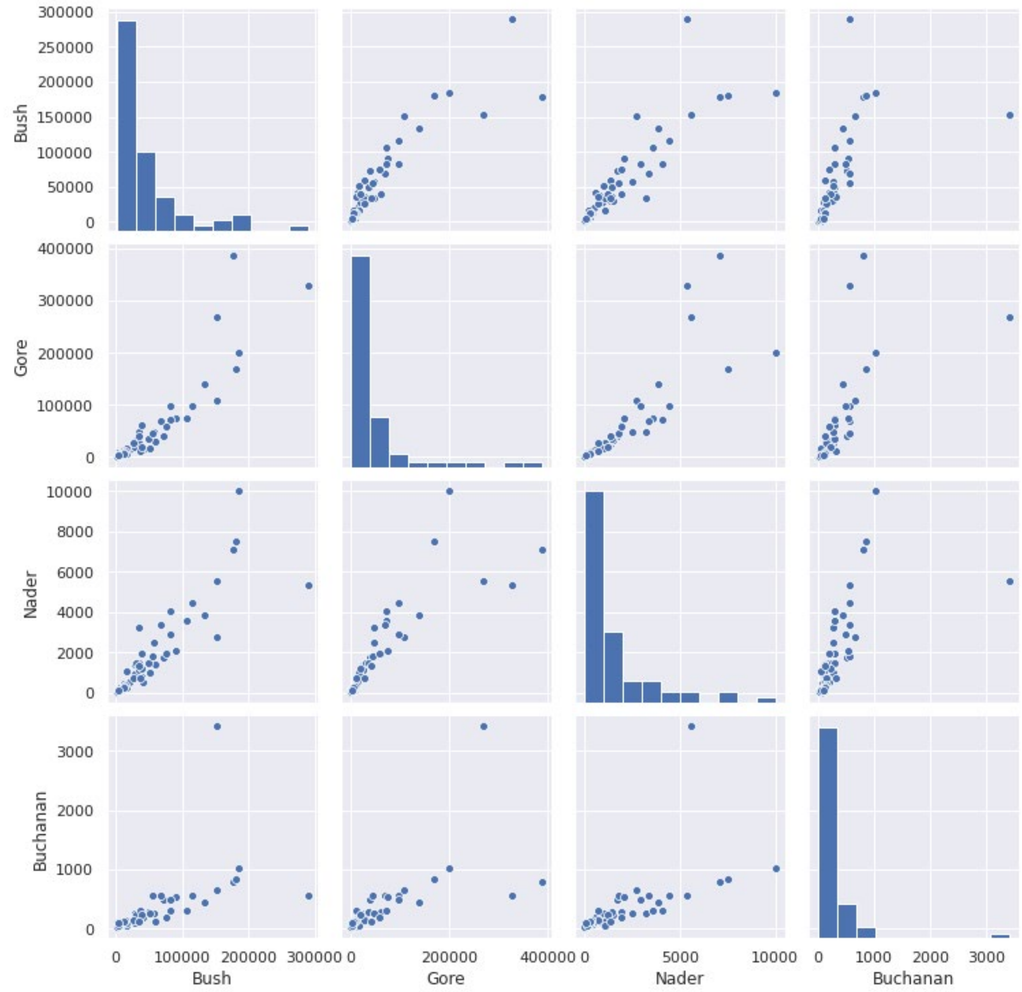
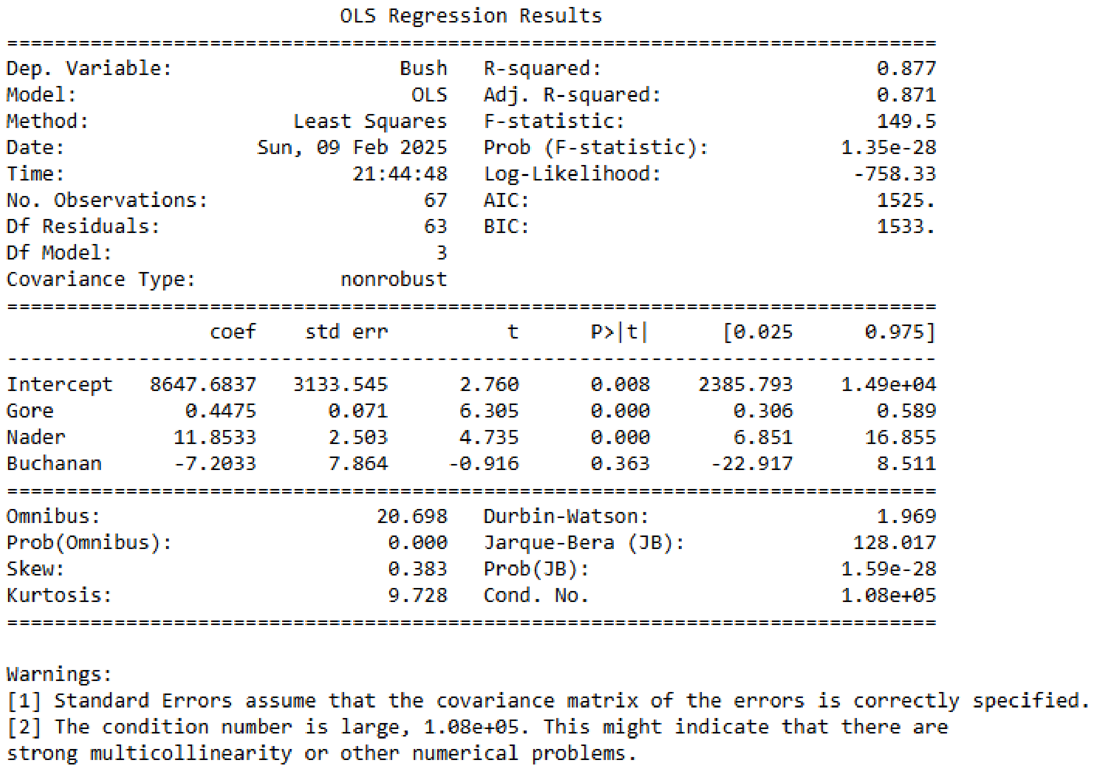
Multiple Linear Regression

The results indicate that while the R-squared value initially increases with higher-degree polynomial terms, it stabilizes at degree 3. This suggests that beyond a certain point, adding more polynomial terms does not significantly enhance the model's explanatory power. Although a third-degree polynomial achieves the highest R-squared, it does not offer a meaningful improvement over a second-degree polynomial. Higher-degree models often lead to overfitting, capturing noise in the data rather than genuine trends, which reduces their generalizability. The minimal increase in R-squared from degree 2 (0.7151) to degree 3 (0.7151) further supports the idea that additional terms do not add substantial predictive value. Moreover, higher-degree polynomials introduce multicollinearity, making the model more sensitive to small changes in input values. The model summaries also indicate that higher polynomial degrees have large p-values, suggesting that these additional terms are not statistically significant.



The pair plot reveals strong positive correlations among the vote counts for Bush, Gore, Nader, and Buchanan. The scatterplots show clear upward trends, particularly between Bush and Gore, as well as between Gore and Nader, indicating that counties with higher votes for one candidate tend to have higher votes for others. This suggests that vote distributions were not independent but followed regional or demographic patterns.

There is also evidence of collinearity, especially between Bush and Gore, as their votes appear to be strongly linked, forming a near-linear pattern. This means that in a regression model, including both variables may introduce multicollinearity, potentially inflating variance and making coefficient estimates less reliable. The histograms along the diagonal further indicate that the data is right-skewed, with a few counties contributing significantly higher vote counts



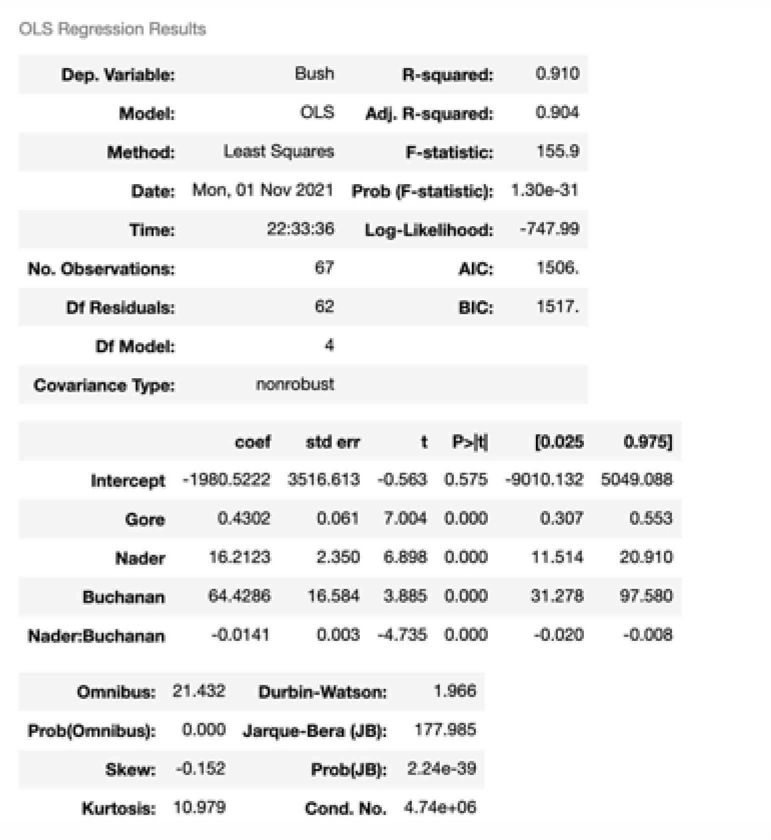
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Bush   R-squared:                       0.877
Model:                            OLS   Adj. R-squared:                  0.871
Method:                 Least Squares   F-statistic:                     149.5
Date:                Sun, 09 Feb 2025   Prob (F-statistic):           1.35e-28
Time:                        21:44:48   Log-Likelihood:                -758.33
No. Observations:                  67   AIC:                             1525.
Df Residuals:                      63   BIC:                             1533.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   8647.6837   3133.545      2.760      0.008    2385.793    1.49e+04
Gore           0.4475      0.071      6.305      0.000       0.306       0.589
Nader         11.8533      2.503      4.735      0.000       6.851      16.855
Buchanan      -7.2033      7.864     -0.916      0.363     -22.917       8.511
==============================================================================
Omnibus:                       20.698   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              128.017
Skew:                           0.383   Prob(JB):                     1.59e-28
Kurtosis:                       9.728   Cond. No.                     1.08e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.08e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The multiple linear regression model predicts the number of votes for Bush using Gore, Nader, and Buchanan as predictors. The R-squared value of 0.877 suggests that the model explains a significant portion of the variance in Bush's vote count. However, the p-values for each predictor reveal that not all features are statistically significant.

The coefficients for Gore ($p < 0.001$) and Nader ($p < 0.001$) are both highly significant, indicating a strong relationship between these predictors and Bush’s vote count. However, the coefficient for Buchanan ($p = 0.363$) is not statistically significant, as its p-value is much higher than the conventional threshold of 0.05. This suggests that Buchanan’s vote count does not contribute meaningfully to predicting Bush’s votes in this model.

Additionally, the large condition number (1.08e+05) indicates possible multicollinearity, meaning that one or more predictor variables may be highly correlated with each other. This could inflate standard errors and make coefficient estimates unreliable. Given these findings, removing Buchanan as a predictor would be an appropriate next step to refine the model.

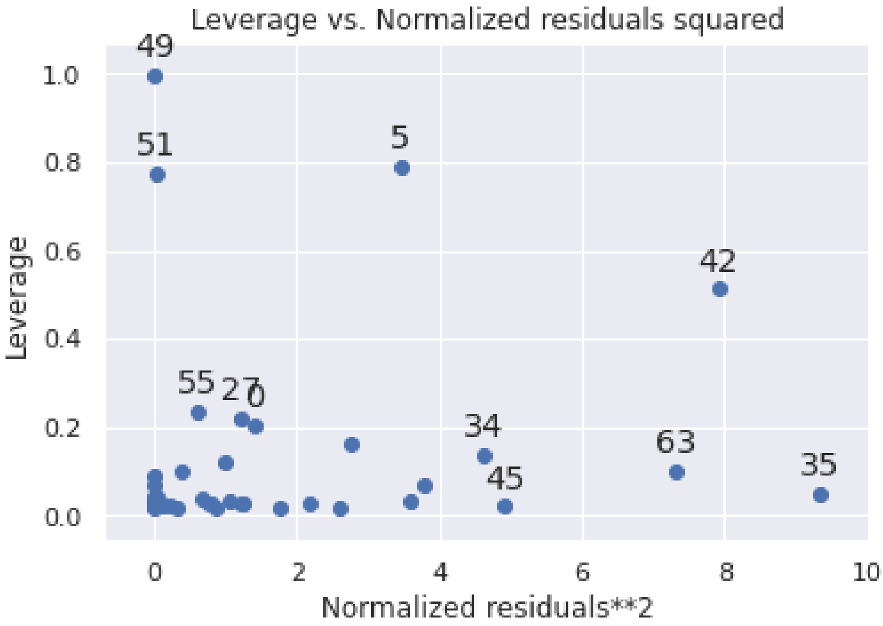


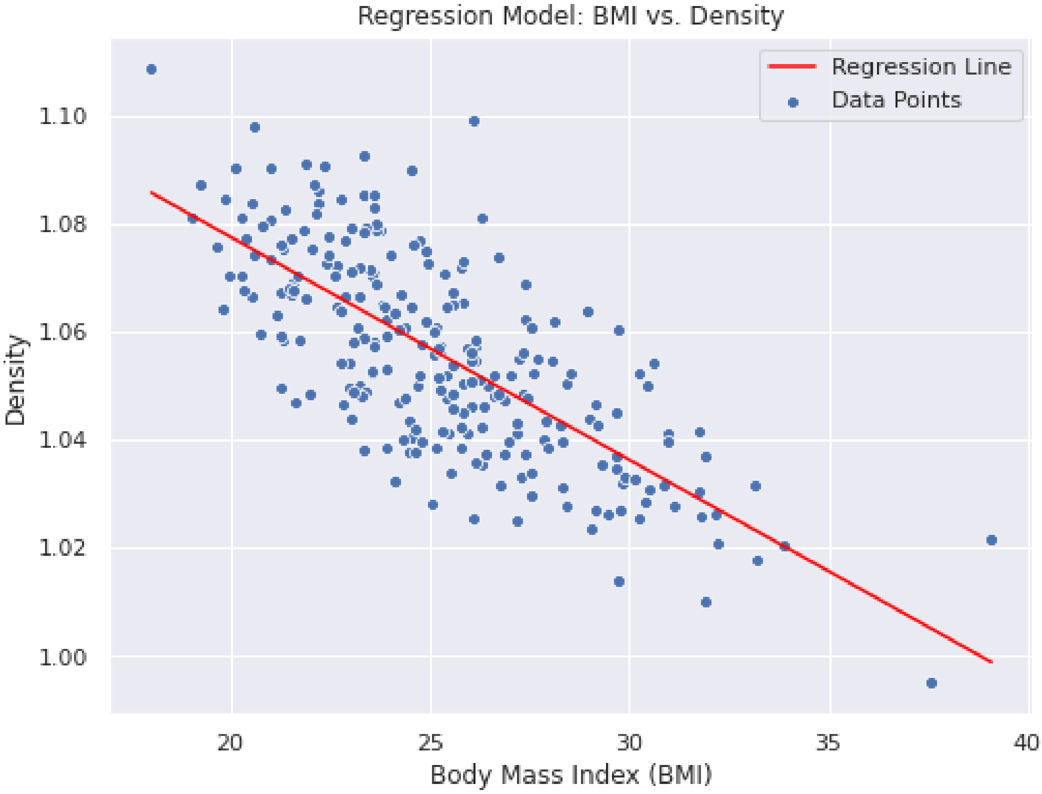
OLS Regression Results

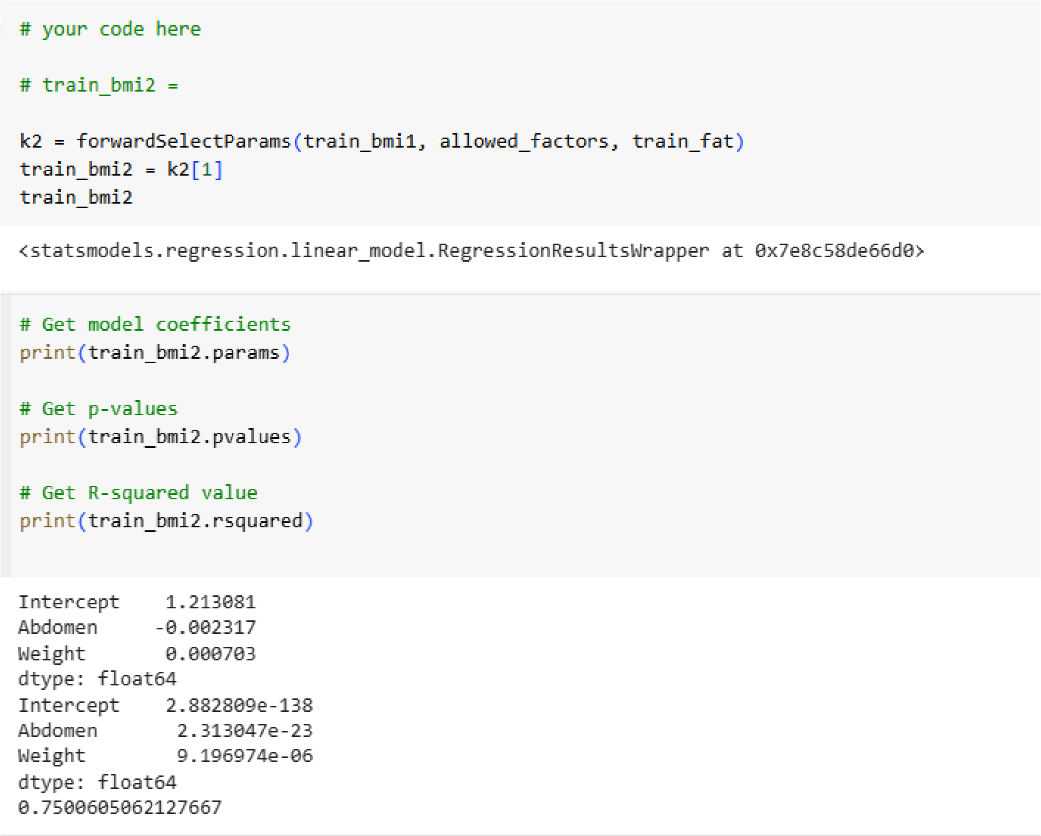
Dep. Variable:	Bush	**R-squared:**	0.910
Model:	OLS	**Adj. R-squared:**	0.904
Method:	Least Squares	**F-statistic:**	155.9
Date:	Mon, 01 Nov 2021	**Prob (F-statistic):**	1.30e-31
Time:	22:33:36	**Log-Likelihood:**	-747.99
No. Observations:	67	**AIC:**	1506.
Df Residuals:	62	**BIC:**	1517.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-1980.5222	3516.613	-0.563	0.575	-9010.132	5049.088
Gore	0.4302	0.061	7.004	0.000	0.307	0.553
Nader	16.2123	2.350	6.898	0.000	11.514	20.910
Buchanan	64.4286	16.584	3.885	0.000	31.278	97.580
Nader:Buchanan	-0.0141	0.003	-4.735	0.000	-0.020	-0.008

Omnibus:	21.432	**Durbin-Watson:**	1.966
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	177.985
Skew:	-0.152	**Prob(JB):**	2.24e-39
Kurtosis:	10.979	**Cond. No.**	4.74e+06


Leverage vs. Normalized residuals squared
Leverage
Normalized residuals**2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
49
51
5
42
55
27
0
34
63
35
45


Regression Model: BMI vs. Density
Regression Line
Data Points
Density
Body Mass Index (BMI)
1.10
1.08
1.06
1.04
1.02
1.00
20
25
30
35
40



```
# your code here

# train_bmi2 =

k2 = forwardSelectParams(train_bmi1, allowed_factors, train_fat)
train_bmi2 = k2[1]
train_bmi2
```

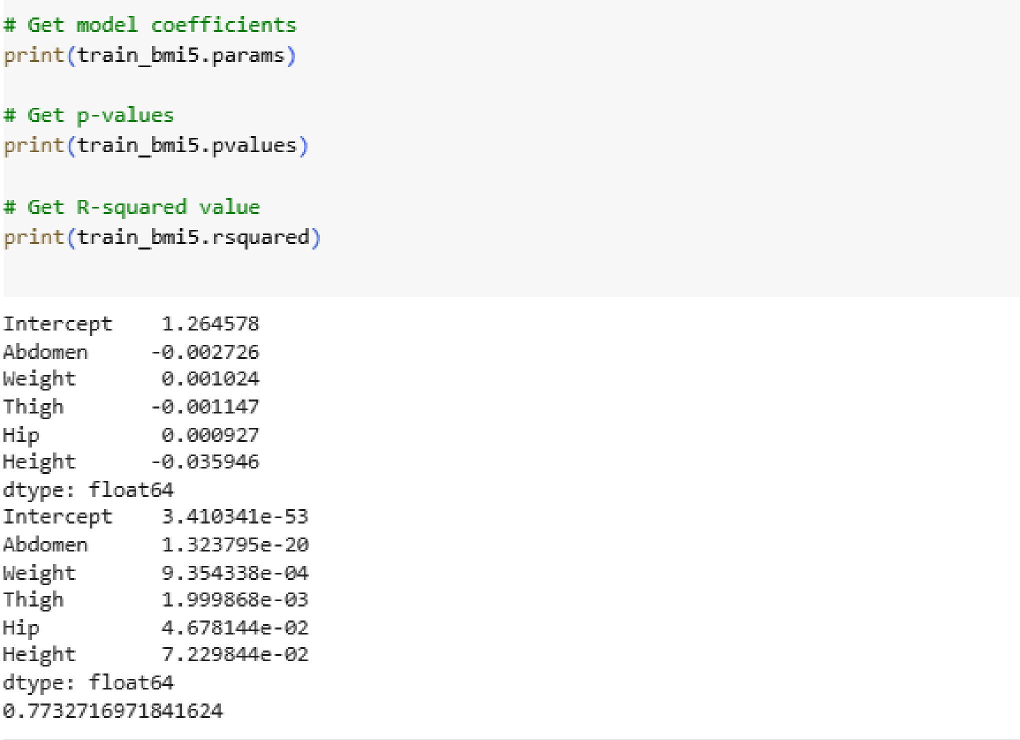
```
<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x7e8c58de66d0>
```

```
# Get model coefficients
print(train_bmi2.params)

# Get p-values
print(train_bmi2.pvalues)

# Get R-squared value
print(train_bmi2.rsquared)
```

```
Intercept    1.213081
Abdomen     -0.002317
Weight       0.000703
dtype: float64
Intercept    2.882809e-138
Abdomen       2.313047e-23
Weight        9.196974e-06
dtype: float64
0.7500605062127667
```

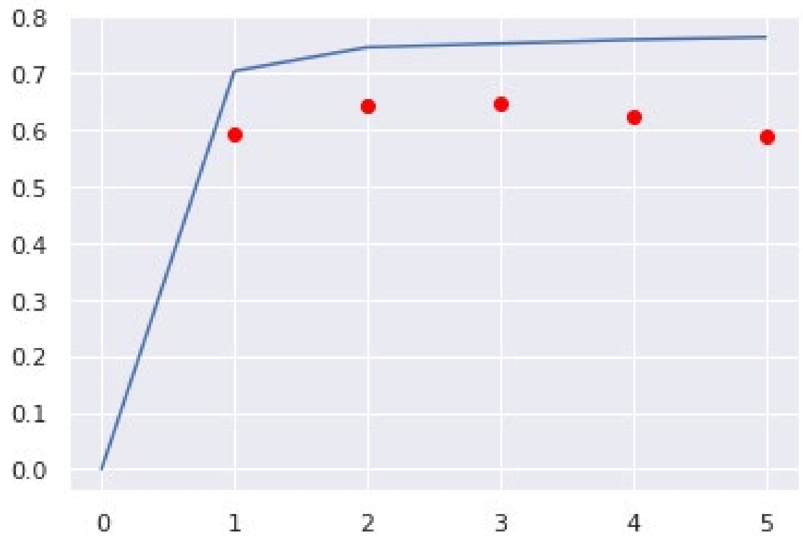


```
# Get model coefficients
print(train_bmi5.params)

# Get p-values
print(train_bmi5.pvalues)

# Get R-squared value
print(train_bmi5.rsquared)
```

```
Intercept    1.264578
Abdomen     -0.002726
Weight       0.001024
Thigh       -0.001147
Hip          0.000927
Height      -0.035946
dtype: float64
Intercept    3.410341e-53
Abdomen      1.323795e-20
Weight       9.354338e-04
Thigh        1.999868e-03
Hip          4.678144e-02
Height       7.229844e-02
dtype: float64
0.7732716971841624
```



Based on the adjusted R-squared vs. number of factors plot, the adjusted R-squared value increases sharply with the first factor and then begins to plateau after approximately two or three factors. This suggests that adding additional factors beyond this point does not significantly improve the model's explanatory power. Since BMI is already a simple model based on height and weight, an enhanced version should aim to balance complexity and predictive strength. The plot indicates that adding more than three factors does not provide substantial gains in adjusted R-squared, making three a reasonable choice for an improved model.

A key consideration is diminishing returns, as the adjusted R-squared increases rapidly at first but stabilizes after three factors, meaning additional factors do not meaningfully enhance predictive ability. Maintaining model simplicity is also important, as a model with fewer factors is easier to interpret and apply in practical settings. Furthermore, including too many factors can lead to overfitting, where the model fits training data well but struggles with new data. Given these considerations, incorporating three factors—height, weight, and one additional factor such as age or body fat percentage—would enhance the BMI model while preserving both simplicity and effectiveness.