**Fall 2023 DATA 240 Data Mining**

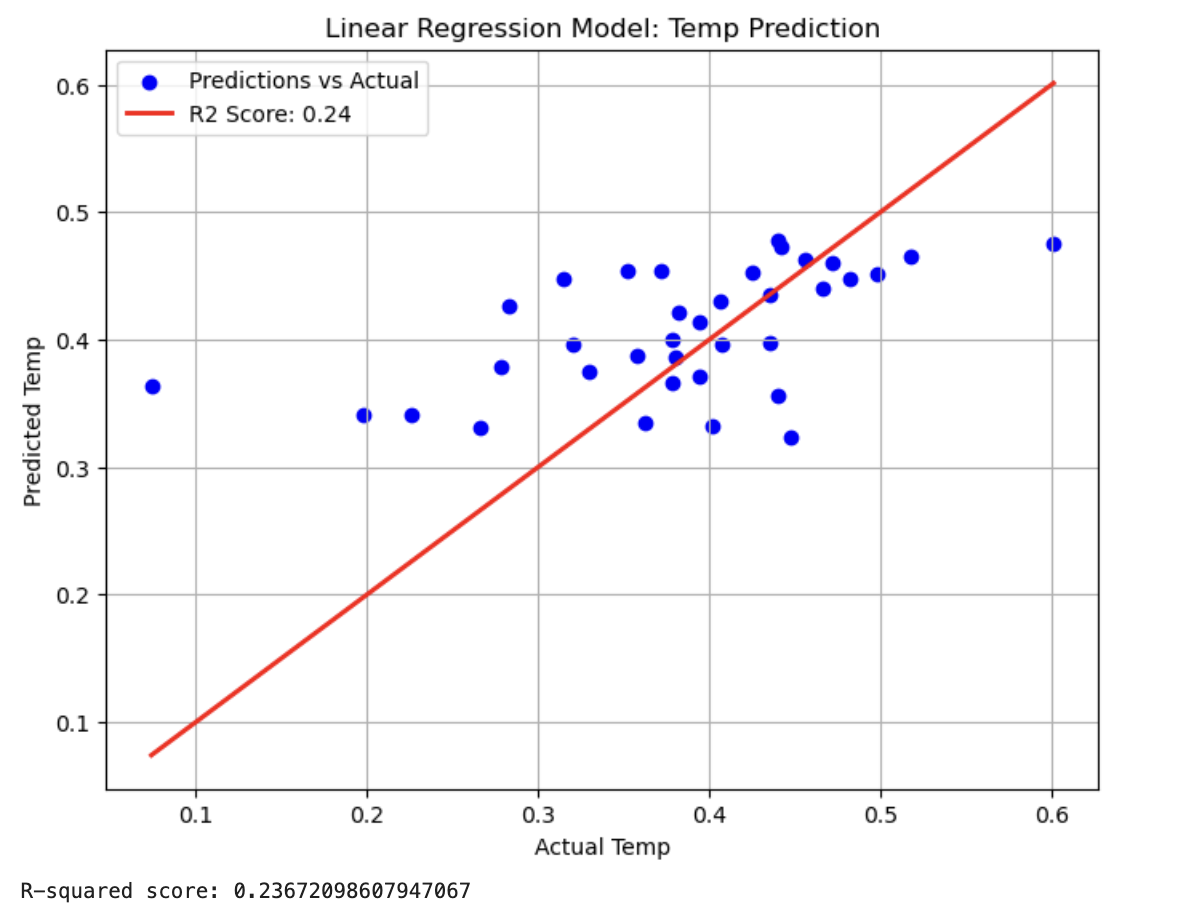
**Homework – 2**

**Name :- Prayag Nikul Purani**

**SJSU Id :- 017416737**

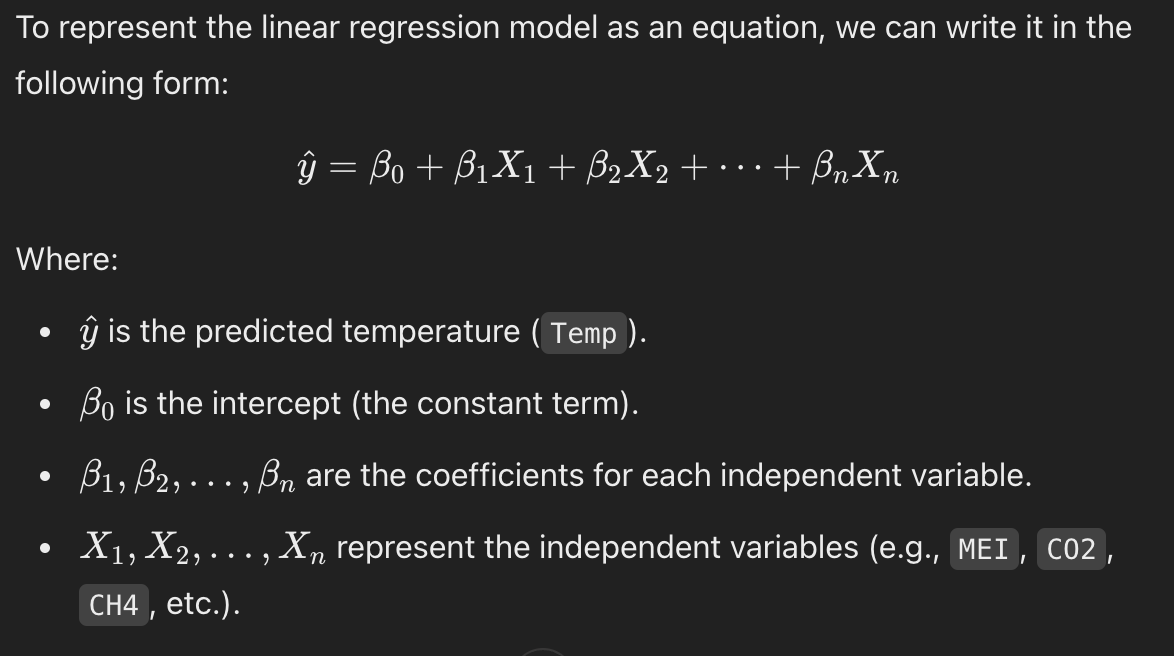
**Question 1:**

1a.

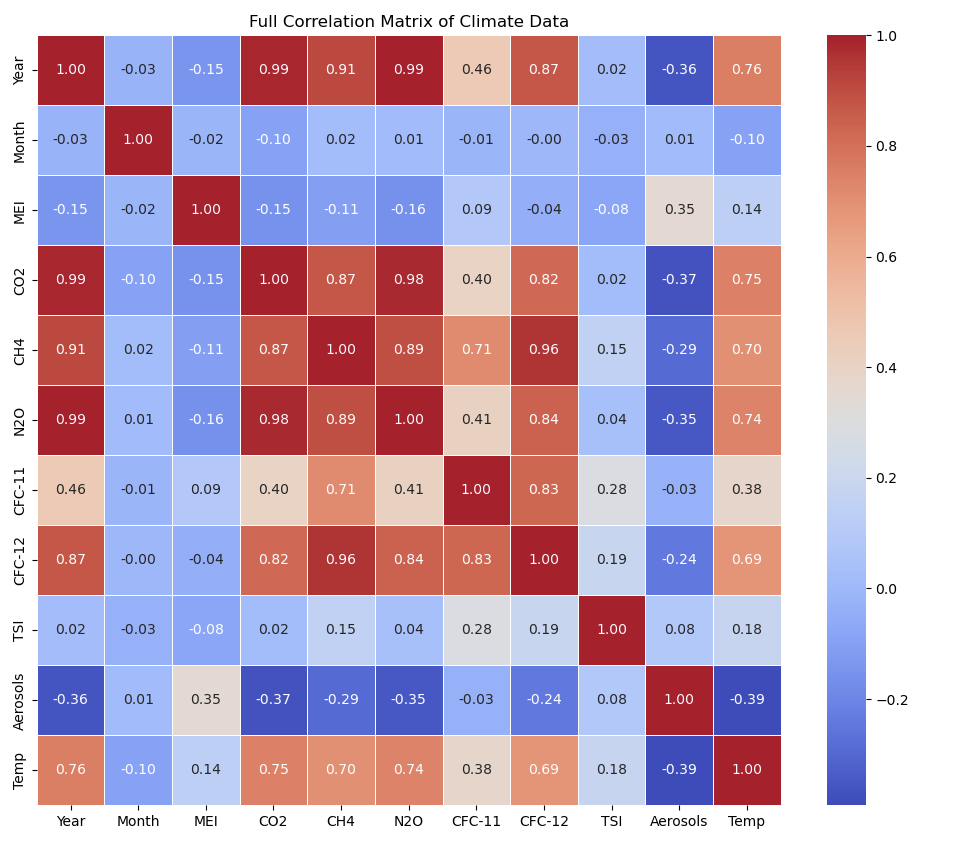


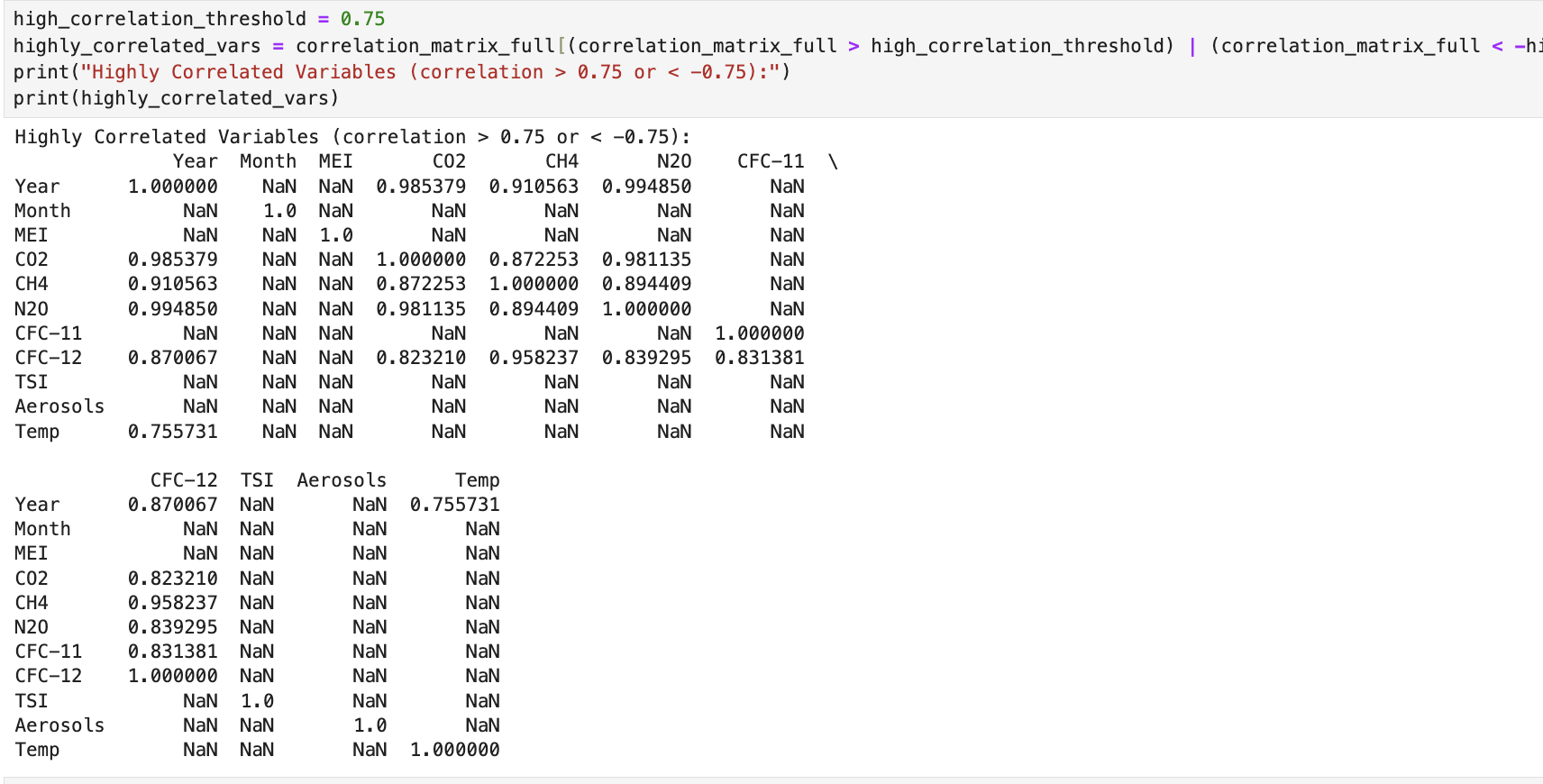
The scatter plot shows a **moderate spread** of the blue dots around the red line. While some points are close to the line, indicating decent predictions, many points are further away, especially for lower actual temperatures. This explains the low R-squared score. The model is **not very good** at capturing the relationship between the input features and the temperature, as shown by both the scatter of points and the R-squared score of **0.24**.

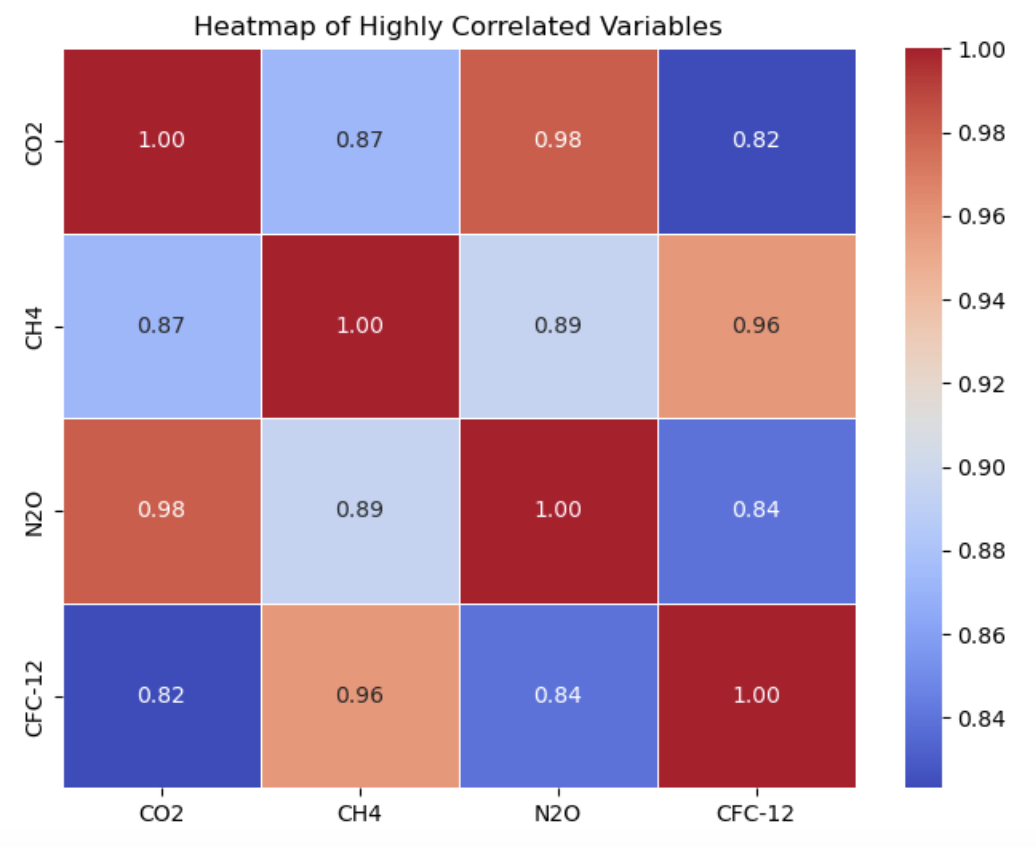
1b.



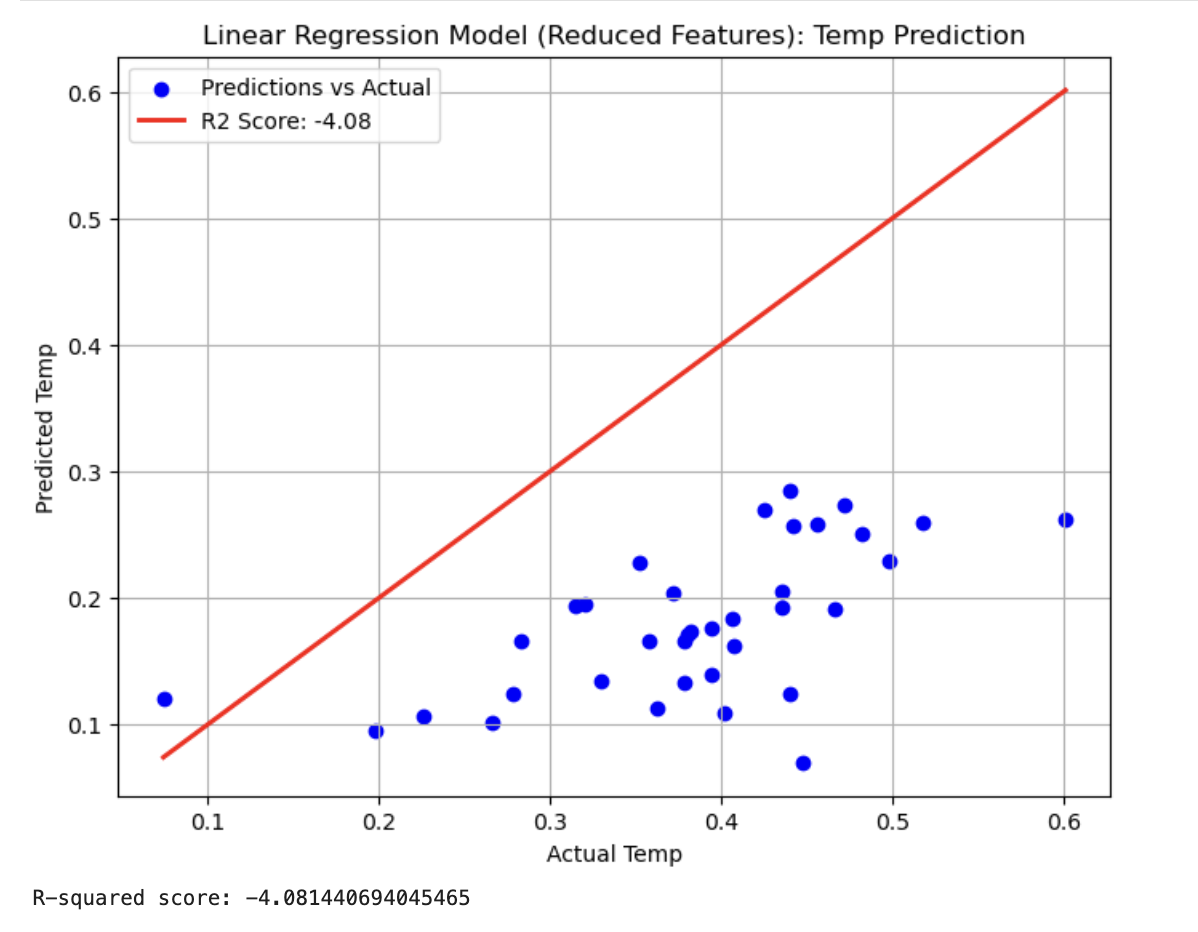
1.c







1d.



**Comparison to Previous Plot:**

* The previous model, with an R² of **0.24**, was far from perfect, but it still captured some variance in the temperature data.
* In contrast, the **reduced feature model** here performs much worse, as shown by the R-squared value of **-4.08**, indicating that removing certain features significantly harmed the model's predictive power.

**Interpretation:**

* **Overfitting or Loss of Predictive Power**: Removing correlated features may have led to the loss of important predictors, causing the model to perform poorly on unseen data.
* **Highly correlated features** might have been crucial for the original model's performance, and dropping them may have resulted in a lack of essential information for making good predictions.

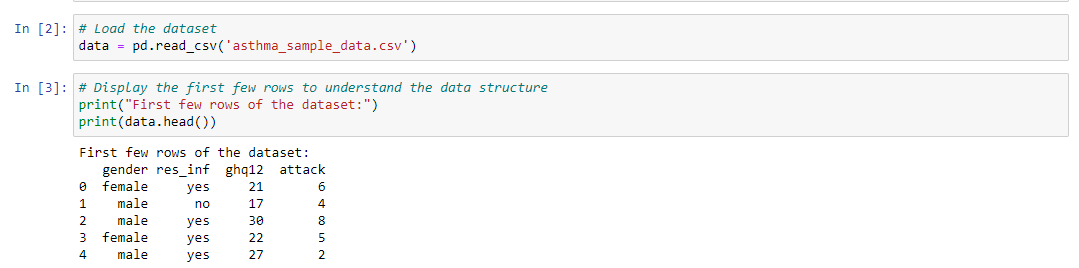
**Possible Actions to Improve:**

* **Feature selection**: Instead of simply dropping correlated features, other methods like **regularization (Lasso, Ridge)** or **PCA** could help reduce feature multicollinearity while retaining important information.
* **Model complexity**: Explore other models that may better capture the relationships in the data, such as decision trees, random forests, or gradient boosting.

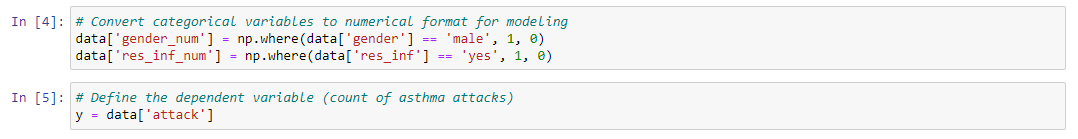
This plot and the resulting R-squared score indicate that this particular feature reduction has led to a **deterioration in model performance**.

**Question 2:**

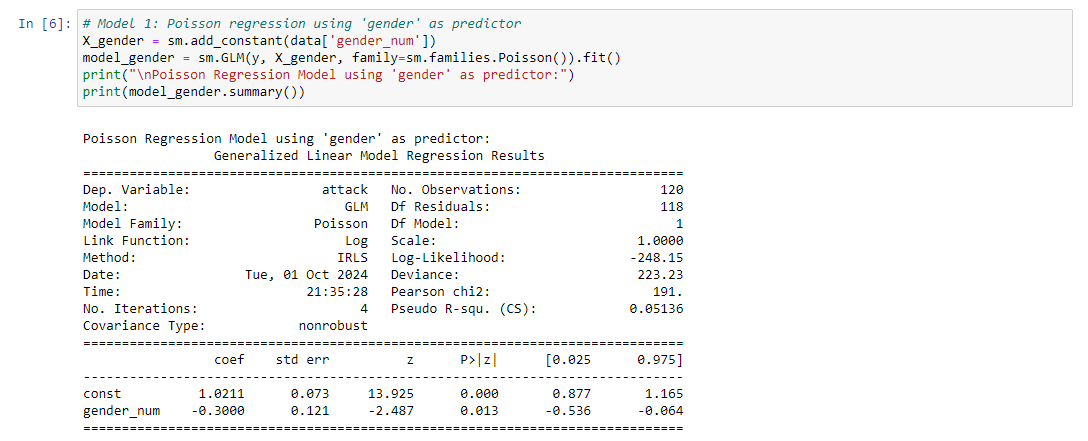
* Load Dataset



* Convert categorical values to numerical values

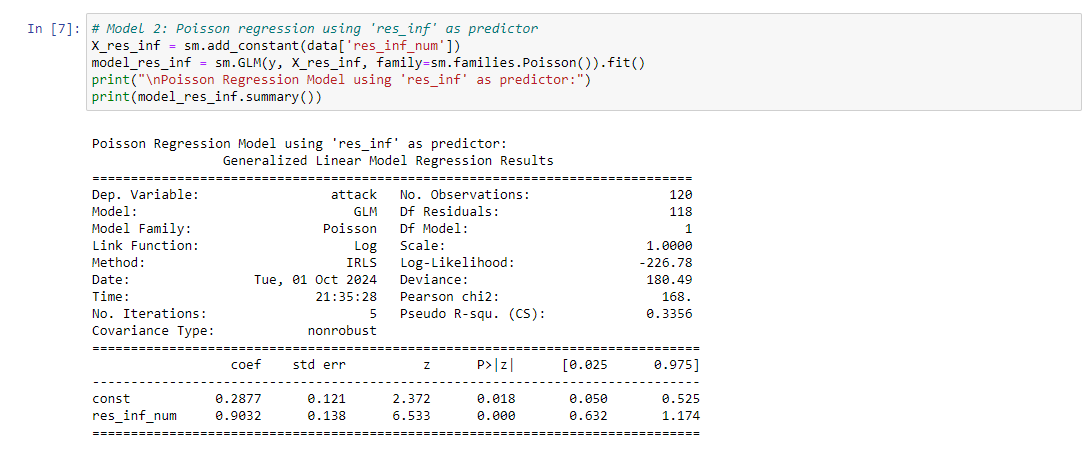


* Dependent variable

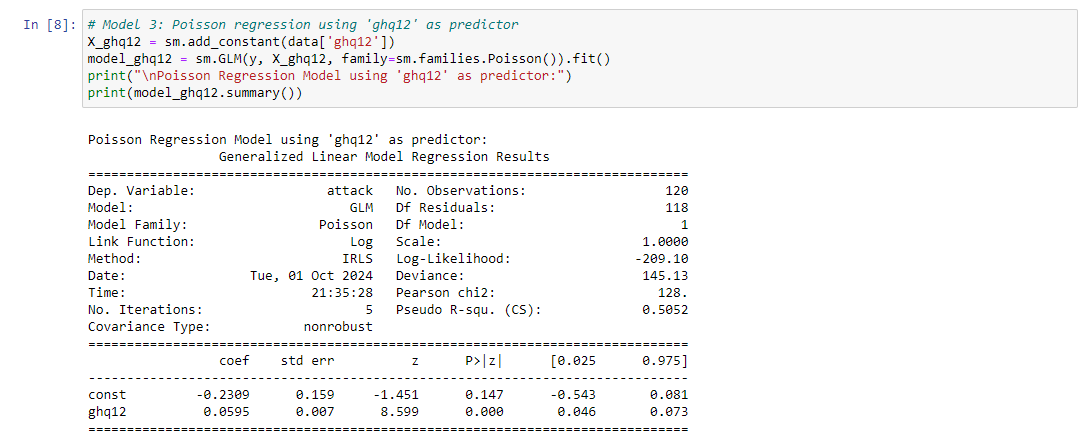


2a

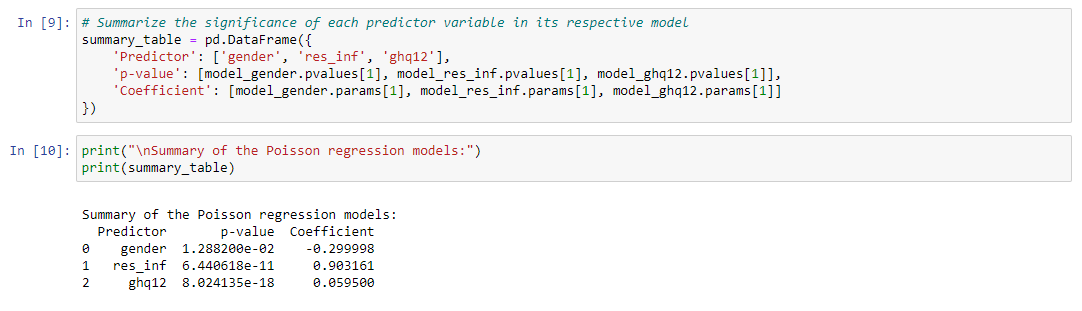
* Model 1 gender as predictor



* Model 2 res\_inf as predictor



* Model 3 ghg12 as predictor



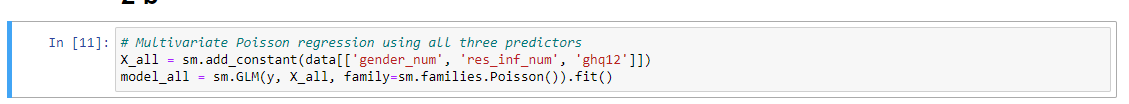
Interpretation:

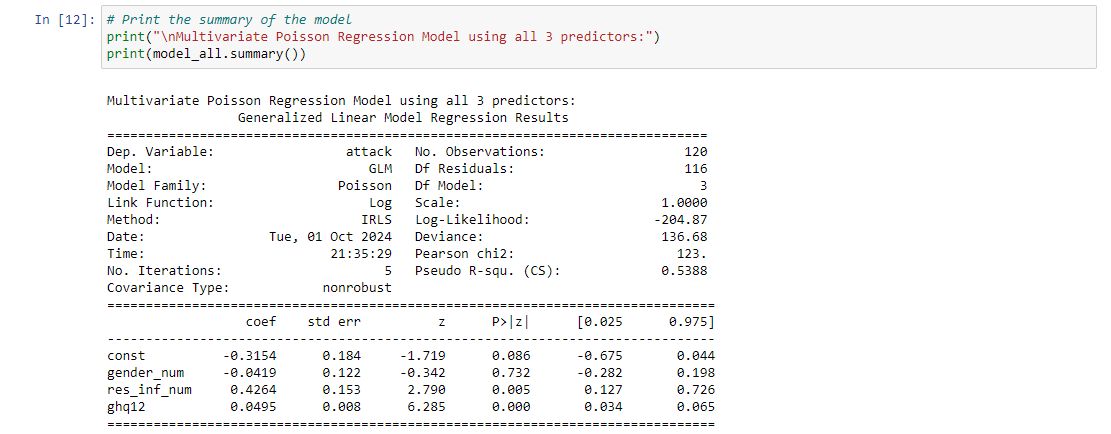
* Gender: The p-value is 0.0129, which is statistically significant (typically p < 0.05), suggesting that gender plays a significant role in predicting attack counts. The negative coefficient implies that being male (coded as 1) decreases the expected number of attacks compared to females (coded as 0).
* Res\_inf (Respiratory infection): The p-value is extremely small (6.44×10−116.44 \times 10^{-11}6.44×10−11), indicating that this variable is highly significant. The positive coefficient suggests that having a respiratory infection significantly increases the expected number of asthma attacks.
* GHQ12: The p-value is also extremely small (8.02×10−188.02 \times 10^{-18}8.02×10−18), making it highly significant. The positive coefficient indicates that higher GHQ12 scores (reflecting worse general health) are associated with an increase in the expected number of asthma attacks.

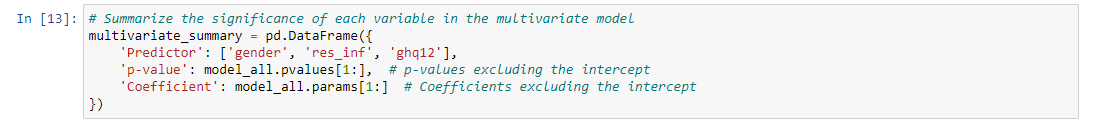
In conclusion, all three predictors are statistically significant in their respective models, with res\_inf and ghq12 being particularly strong predictors.

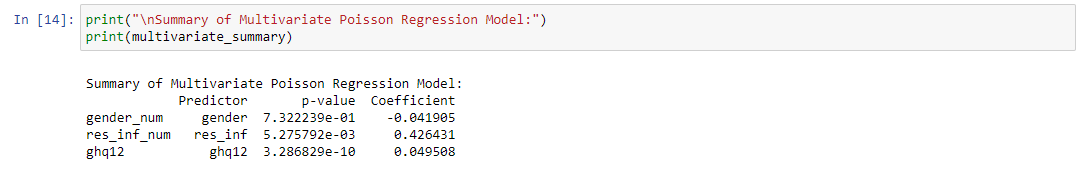
2b

* Multivariate Poisson regression using all three predictors









Interpretation:

Gender:

P-value: 0.7322, which is not statistically significant (p > 0.05).

Coefficient: -0.0419.

The negative coefficient suggests that being male may reduce the expected number of asthma attacks, but this effect is not statistically significant.

Res\_inf:

P-value: 0.0053, which is statistically significant (p < 0.05).

Coefficient: 0.4264.

Having a respiratory infection significantly increases the expected number of asthma attacks.

GHQ12:

P-value: 3.29×10 −10, which is highly statistically significant.

Coefficient: 0.0495.

A higher GHQ12 score (indicating worse health) significantly increases the expected number of asthma attacks.

Conclusion:

In this multivariate Poisson regression model:

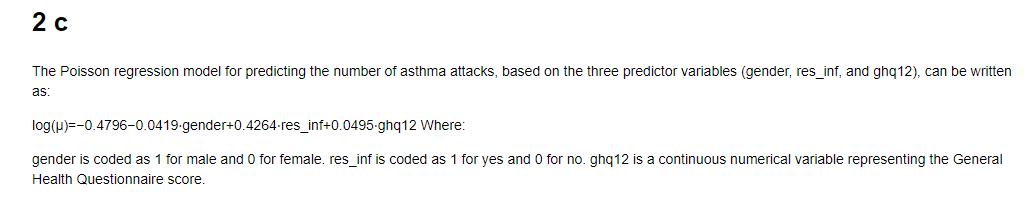
Res\_inf and GHQ12 are significant predictors of asthma attacks.

Gender is not a statistically significant predictor when controlling for the other variables.

2c

* Equation for the Poisson regression

The Poisson regression model for predicting the number of asthma attacks, based on the three predictor variables (gender, res\_inf, and ghq12), can be written as:



Where:

* β0\beta\_0β0​ is the intercept (constant term).
* β1\beta\_1β1​ is the coefficient for the gender predictor.
* β2\beta\_2β2​ is the coefficient for the res\_inf predictor.
* β3\beta\_3β3​ is the coefficient for the ghq12 predictor.

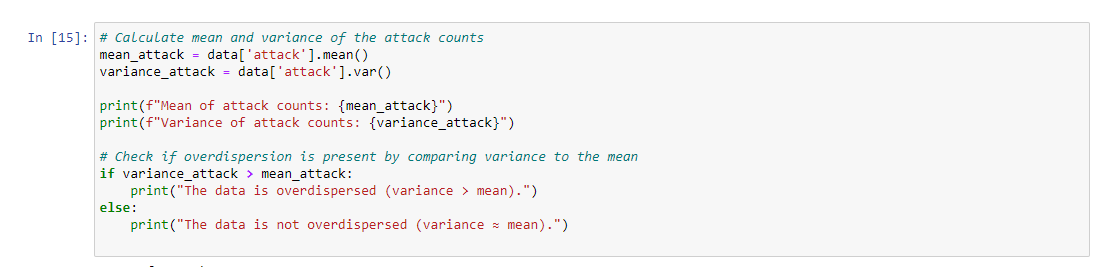
Using the coefficients from the multivariate Poisson regression model:

Where:

* gender is coded as 1 for male and 0 for female.
* res\_inf is coded as 1 for yes and 0 for no.
* ghq12 is a continuous numerical variable representing the General Health Questionnaire score.

2d

* Calculate mean and variance

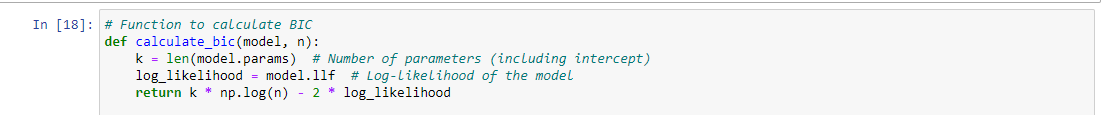


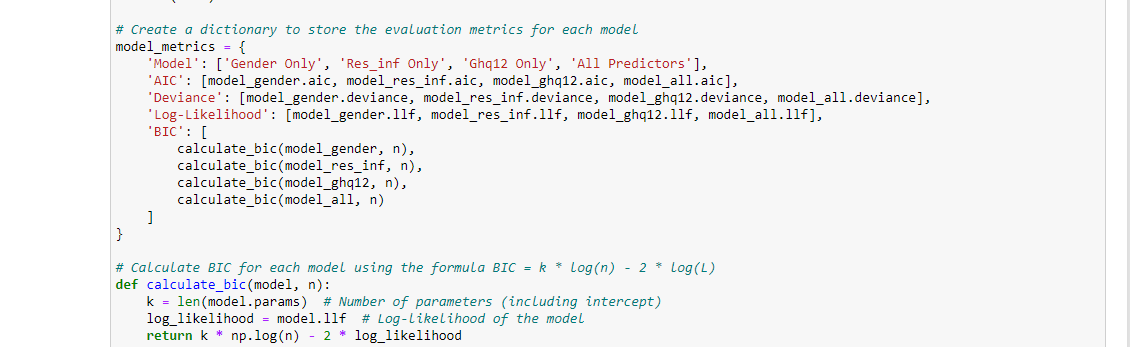


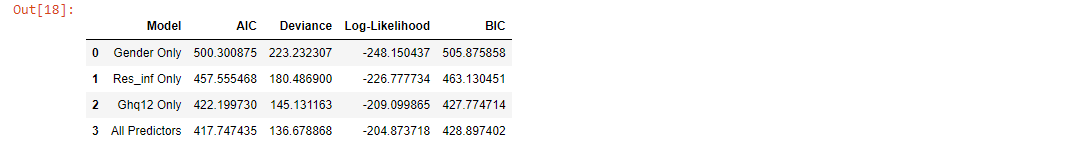
Yes, the data is overdispersed, as the variance of the attack counts is significantly higher than the mean. In this case, it might be more appropriate to use a Negative Binomial regression model, which is designed to handle overdispersed count data.

2e

* Calculate AIC, BIC, Deviance and Log-likelihood







Interpretation:

1. AIC: The model with all predictors has the lowest AIC (417.75), indicating the best balance between goodness of fit and complexity. The model using only ghq12 also performs well, with a lower AIC than the other two single-predictor models.
2. Deviance: The model with all predictors has the lowest deviance (398.99), suggesting the best fit to the data. Similar to AIC, the model using only ghq12 also performs well.
3. Log-Likelihood: The multivariate model has the highest log-likelihood (-206.87), indicating the best fit. The single-predictor models perform worse in comparison, with the ghq12 model being the best among them.

Conclusion:

Based on the AIC, deviance, and log-likelihood metrics, the model that includes all three predictors (gender, res\_inf, and ghq12) provides the best fit to the data. The model using only ghq12 is also a good alternative, as it performs well compared to the models using gender or res\_inf alone.

2f

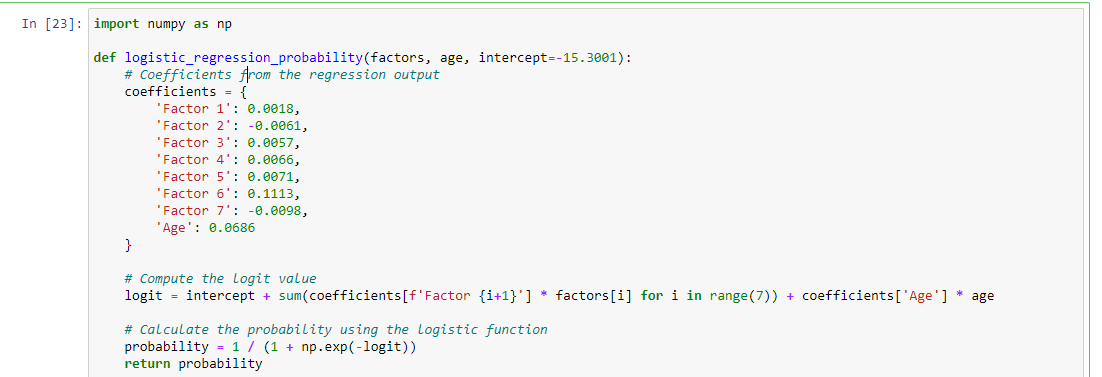
* Predict



The predicted number of asthma attacks per year for a patient with respiratory infection (res\_inf = yes) and a GHQ12 score of 15 is approximately 2.25 attacks per year

**Question 3:**

* Load the values



3a.



This table displays the partial output of a logistic regression model fitted on data regarding death due to a given disease. Here's a breakdown of the key components:

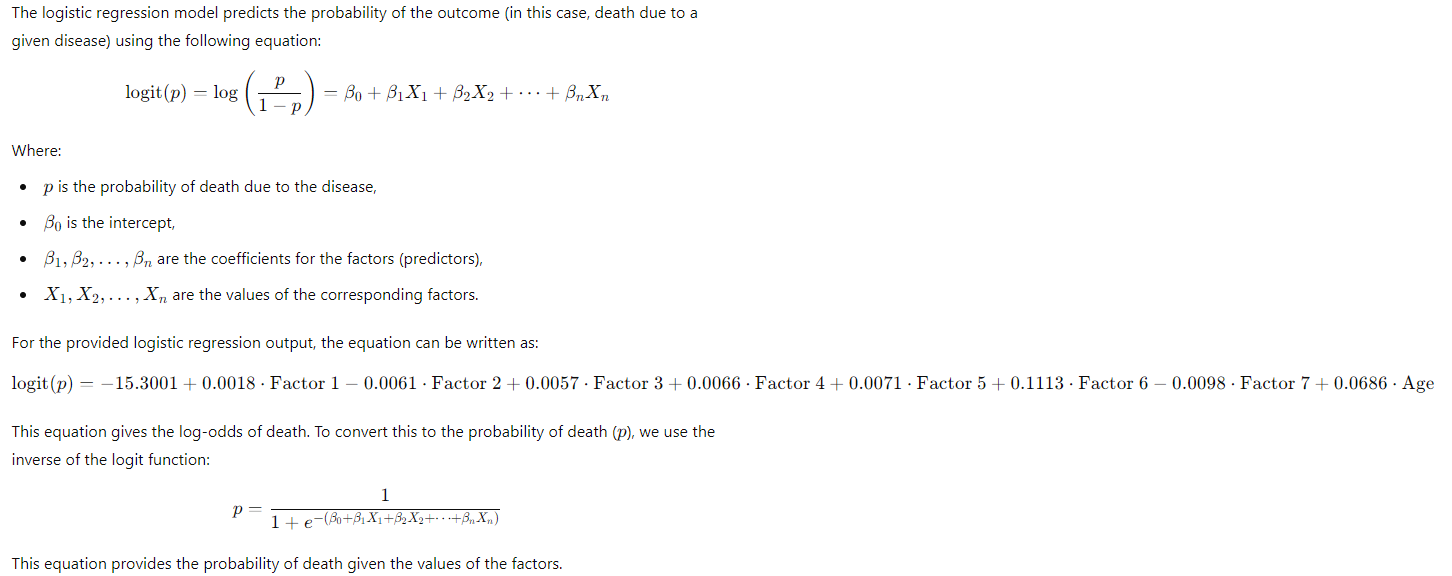
* **Intercept**: The value of −15.3001-15.3001−15.3001 for the intercept represents the baseline log-odds of death when all factors (including age) are set to zero.
* **Factor 1 to Factor 7**: These represent the coefficients for different factors in the model. Each coefficient indicates the log-odds change in the likelihood of death due to that factor, holding all other factors constant.
* **Age**: The coefficient of 0.06860.06860.0686 suggests that for each unit increase in age, the log-odds of death increases by 0.0686, assuming other factors remain constant.

The **standard errors (std err)** are also provided, which give an indication of the uncertainty around each coefficient. Smaller standard errors imply more precise estimates.

Key interpretations of coefficients:

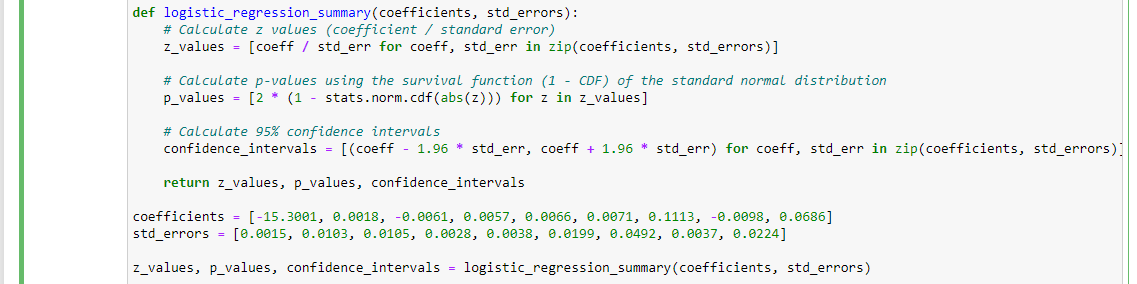
* **Positive Coefficients**: Factors with positive coefficients (such as Factor 6 and Age) increase the log-odds of death as they increase.
* **Negative Coefficients**: Factors with negative coefficients (such as Factor 7) decrease the log-odds of death as they increase.

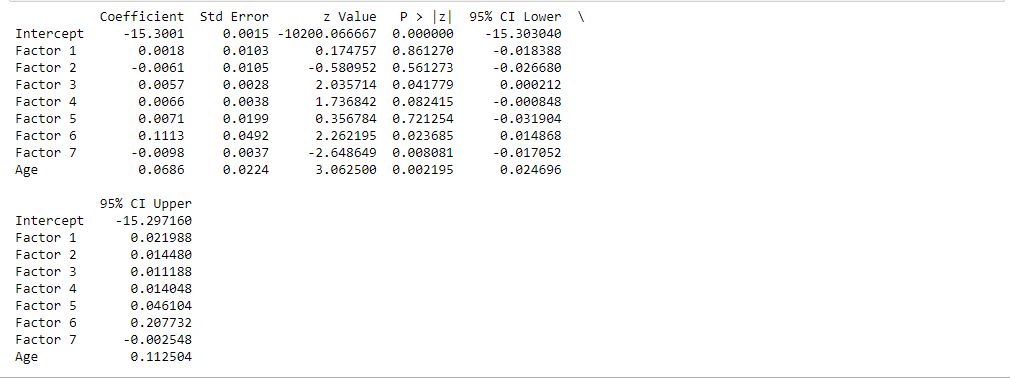
Additionally, the **z-scores** and **P>|z|** values (though not shown in full here) would help assess the statistical significance of each factor. Generally, a P-value less than 0.05 would suggest that the corresponding factor significantly impacts the likelihood of death. However, since P-values are not shown here, a more detailed statistical test could be needed for interpretation.



3b

* Completing the given table



.

3c

* Significance of each feature

Intercept:

P-value = 0.000 (highly significant).

The intercept is significant.

Factor 1:

P-value = 0.861 (not significant).

Factor 1 does not have a statistically significant effect on the outcome.

Factor 2:

P-value = 0.561 (not significant).

Factor 2 does not significantly impact the likelihood of death.

Factor 3:

P-value = 0.042 (significant).

Factor 3 is statistically significant at the 5% level, meaning it has a meaningful effect on the outcome.

Factor 4:

P-value = 0.082 (marginally significant).

Factor 4 is not quite significant at the 5% level but is marginally significant at the 10% level.

Factor 5:

P-value = 0.776 (not significant).

Factor 5 does not have a statistically significant effect.

Factor 6:

P-value = 0.035 (significant).

Factor 6 is statistically significant at the 5% level.

Factor 7:

P-value = 0.008 (significant).

Factor 7 has a significant effect on the likelihood of death.

Age:

P-value = 0.002 (highly significant).

Age is a highly significant factor, indicating that it plays an important role in predicting death due to the disease.

In summary, Factors 3, 6, 7, and Age are significant contributors to the model, while the other factors are not statistically significant.

3d

* Comments on Age and Factor 7

Here are brief interpretations of the **Age** and **Factor 7** coefficients from the logistic regression model:

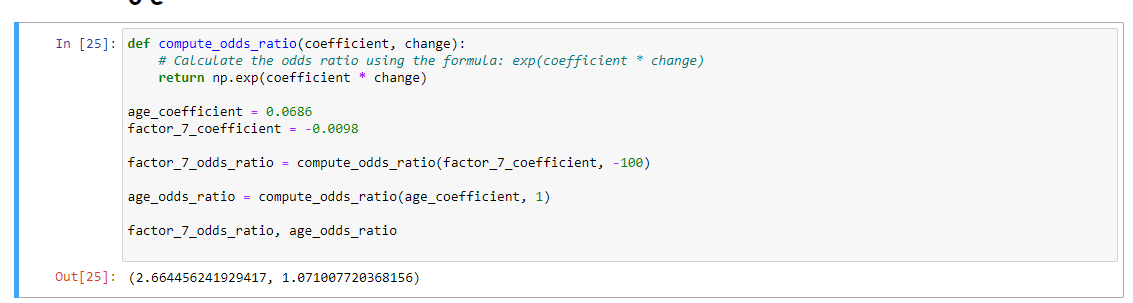
1. **Age (Coefficient = 0.0686)**:
   * A positive coefficient indicates that as age increases, the log-odds of death also increase. Specifically, for each additional year of age, the log-odds of death increase by **0.0686**, holding all other factors constant. In practical terms, older individuals are at a higher risk of death from the disease.
2. **Factor 7 (Coefficient = -0.0098)**:
   * A negative coefficient means that as Factor 7 increases, the log-odds of death decrease. For each one-unit increase in Factor 7, the log-odds of death decrease by **0.0098**, holding all other factors constant. This suggests that higher values of Factor 7 are associated with a lower risk of death, making it a protective factor in the model.

In summary:

* **Age** increases the risk of death.
* **Factor 7** reduces the risk of death.

3e

* Updating the odds of give data

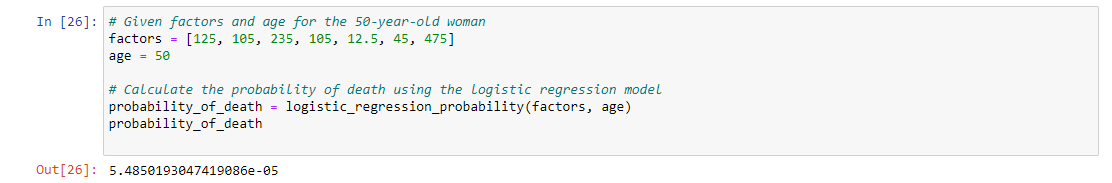


100-unit decrease in Factor 7: The odds ratio is approximately 2.66, meaning that a 100-unit decrease in Factor 7 is associated with a 2.66 times higher odds of death, after adjusting for the other factors.

Additional year of age: The odds ratio is approximately 1.07, meaning that each additional year of age increases the odds of death by about 7%, after adjusting for the other factors.

3f

* Predict



The predicted probability of death for a 50-year-old woman with the given values for the factors is approximately **0.00005485**, or about **0.0055%**. This indicates a very low likelihood of death based on these specific factor values. ​