Homework 6 - Case Study 3

Stephen Wojcik

2025-07-26

What Data I Collected

For Case Study 6 I found a data set with data corresponding to prostate cancer risk levels in humans. This data contained factors such as amount of sleep, age, BMI as well as a risk factor level of developing prostate cancer.

Why This Top Is Interesting and Important

This topic is interesting to me as I have had family members who have had prostate cancer. The data presents real life factors that you can control. By knowing what aspects can reduce your risks you can make the choices to manage your health in positive ways. It was rewarding to work on a project that shed light on an important subject to me.

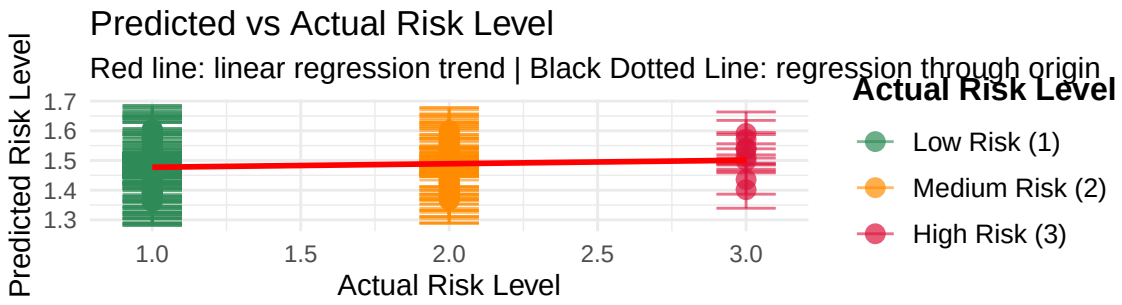
How Did I Analyze The data

I analyzed the data by using linear regression. Linear regression is a data analysis technique that uses independent and dependent variables to create lines that best match the data. These lines are able to give an understanding of the variables selected. Additionally, linear regression is able to have multiple dependent variables for one independent variable by expanding the equation $y = mx + b$. This allows for the understanding of how multiple factors can lead to a specific outcome.

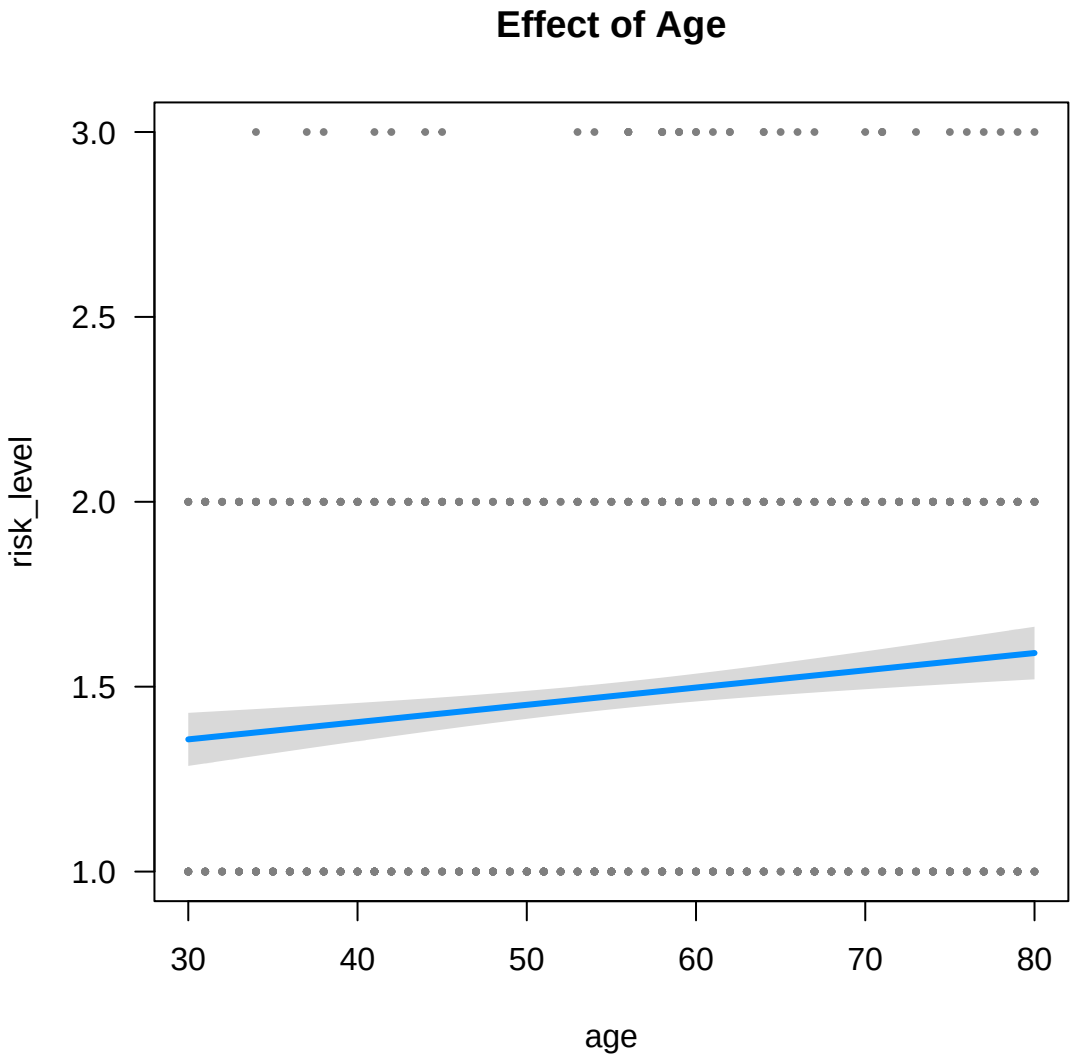
What Did I Find In The Data

By using linear regression on the data set I was able to find some interesting connections. The first predictor that I looked at while reviewing the data was age vs risk level. I was able to get a good Prediction vs Actual rating for these data points. This proved to be a reliable indicator because the data points are tightly packed around the redline. I continued evaluating the data between the many variables.

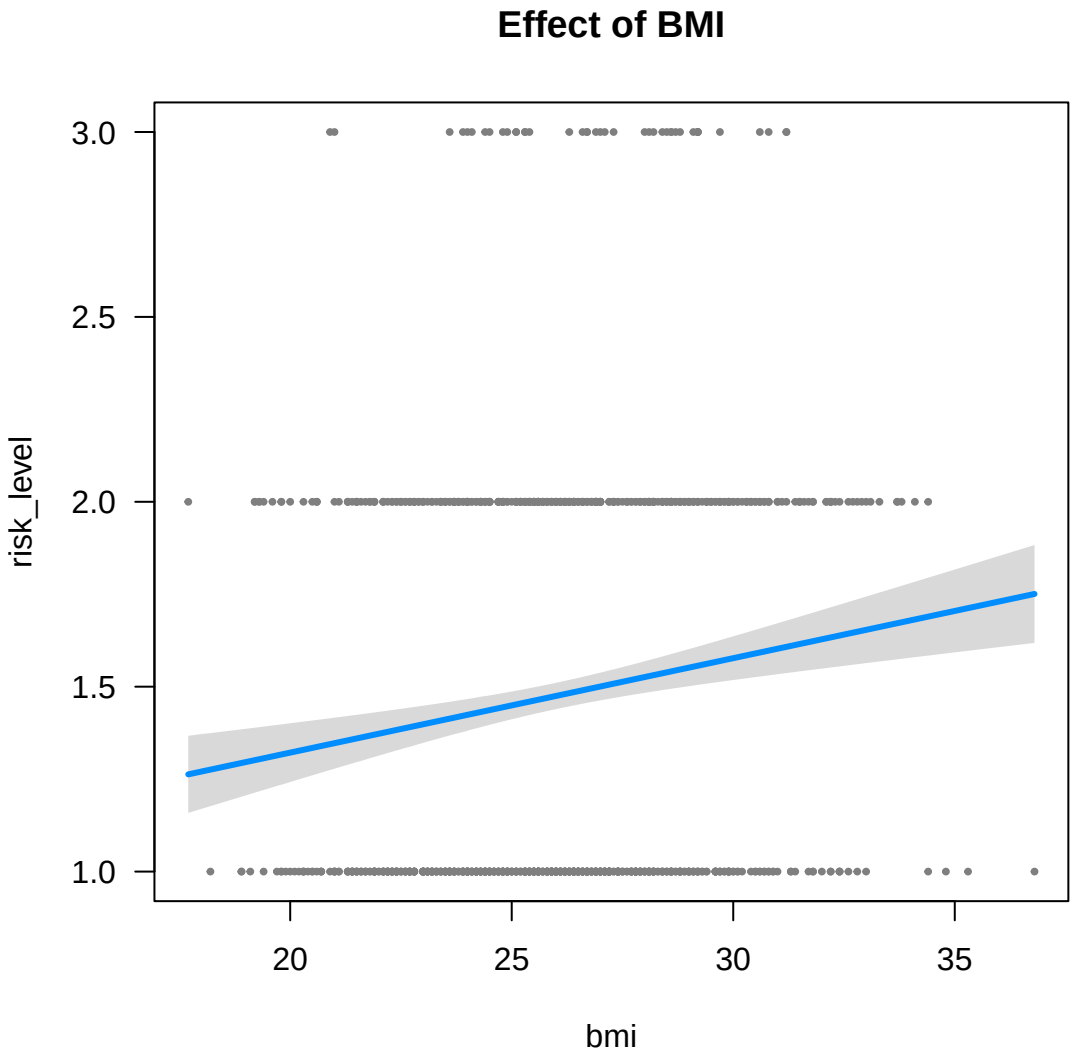
In the case of age, the data showed that with the advancement of age the risk of developing prostate cancer increased. As you can see from the Predicted vs Actual Risk Level Graph the data was tightly packed around the red line which is a indicator that our line has a high degree of confidence.



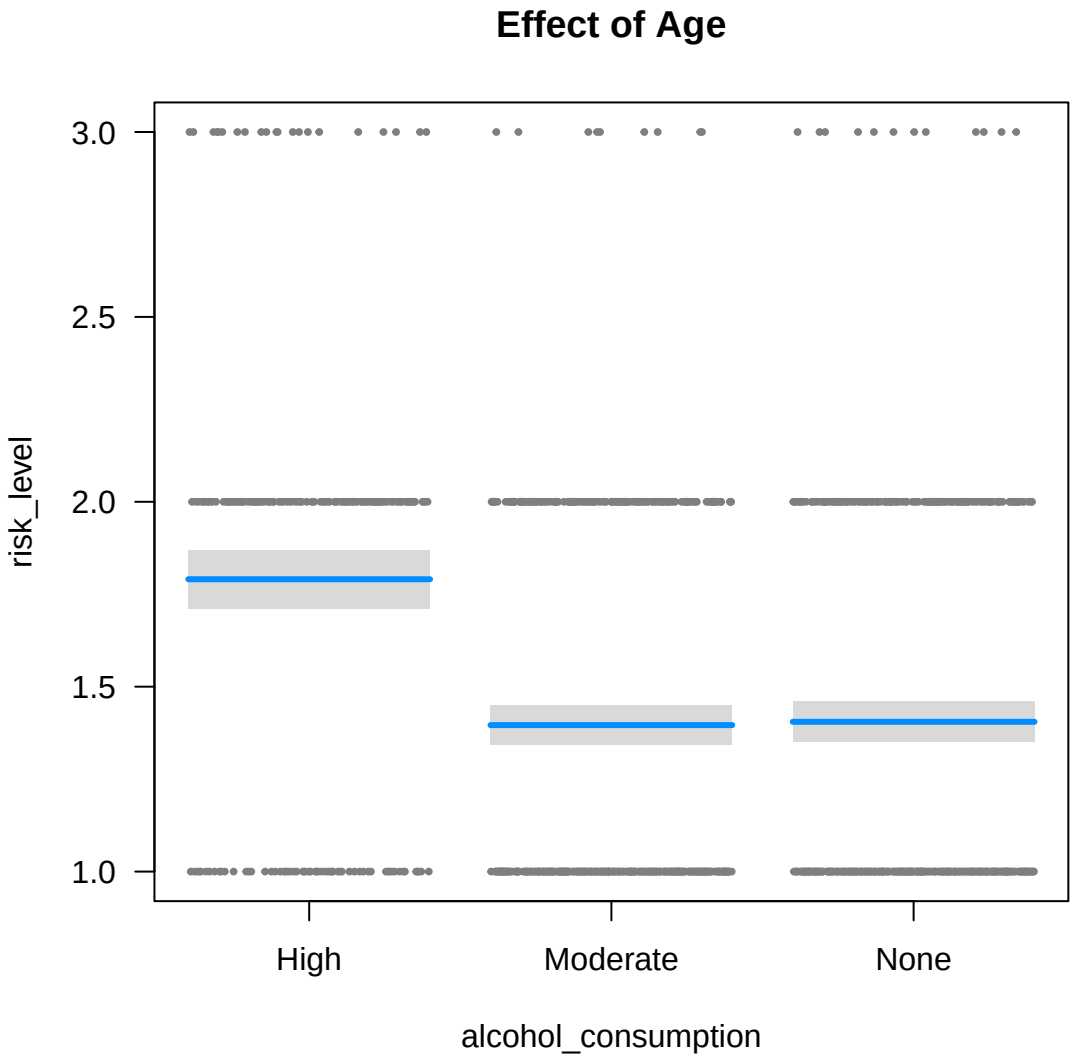
As we can see from the regression visualization for age verse risk level as age increases the risk factor increases.



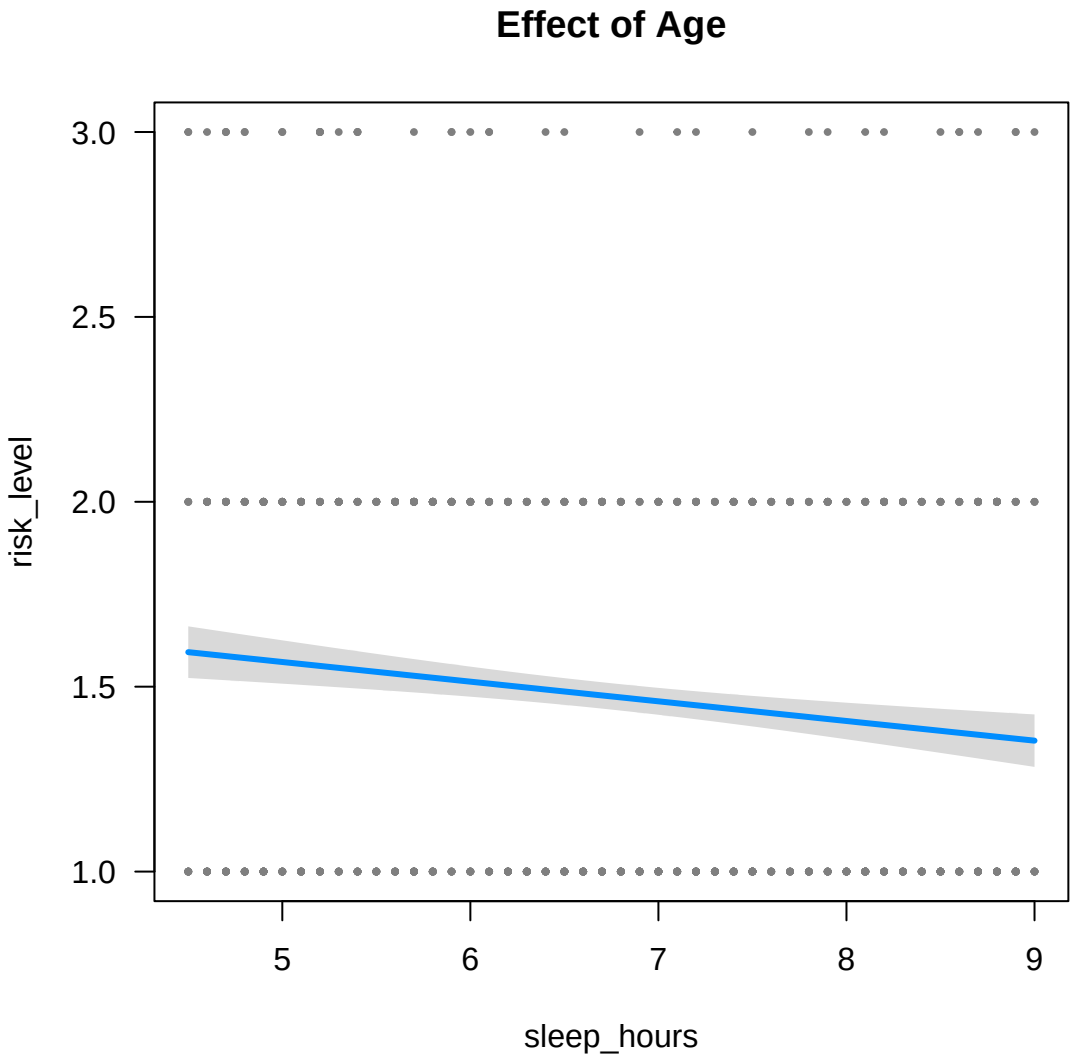
Additionally we can see a similar trend in BMI as age. The higher your BMI the higher your risk for colon cancer.



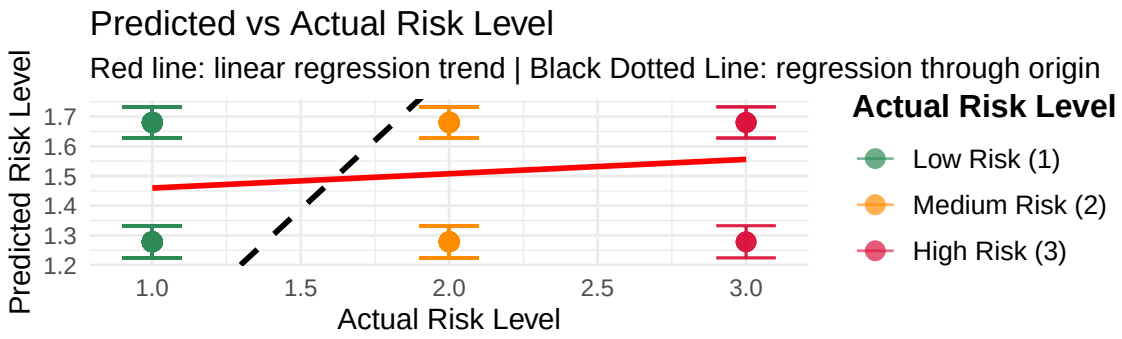
The results on the consumption of alcohol showed that a heavy consumption increased your risk. Moderate and no alcohol consumption increased the risk less.



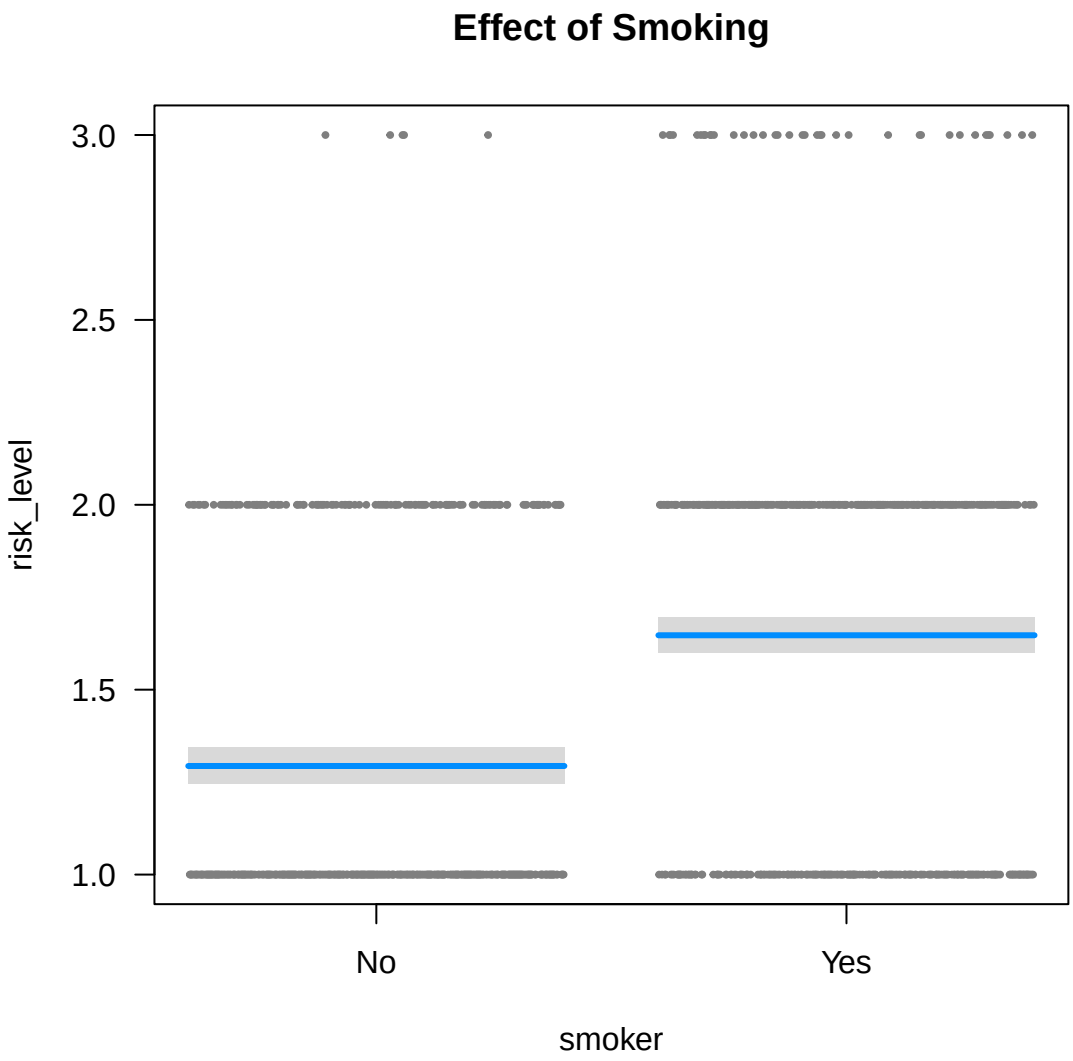
The data showed that sleep was an important factor in lowering your risk of colon cancer.



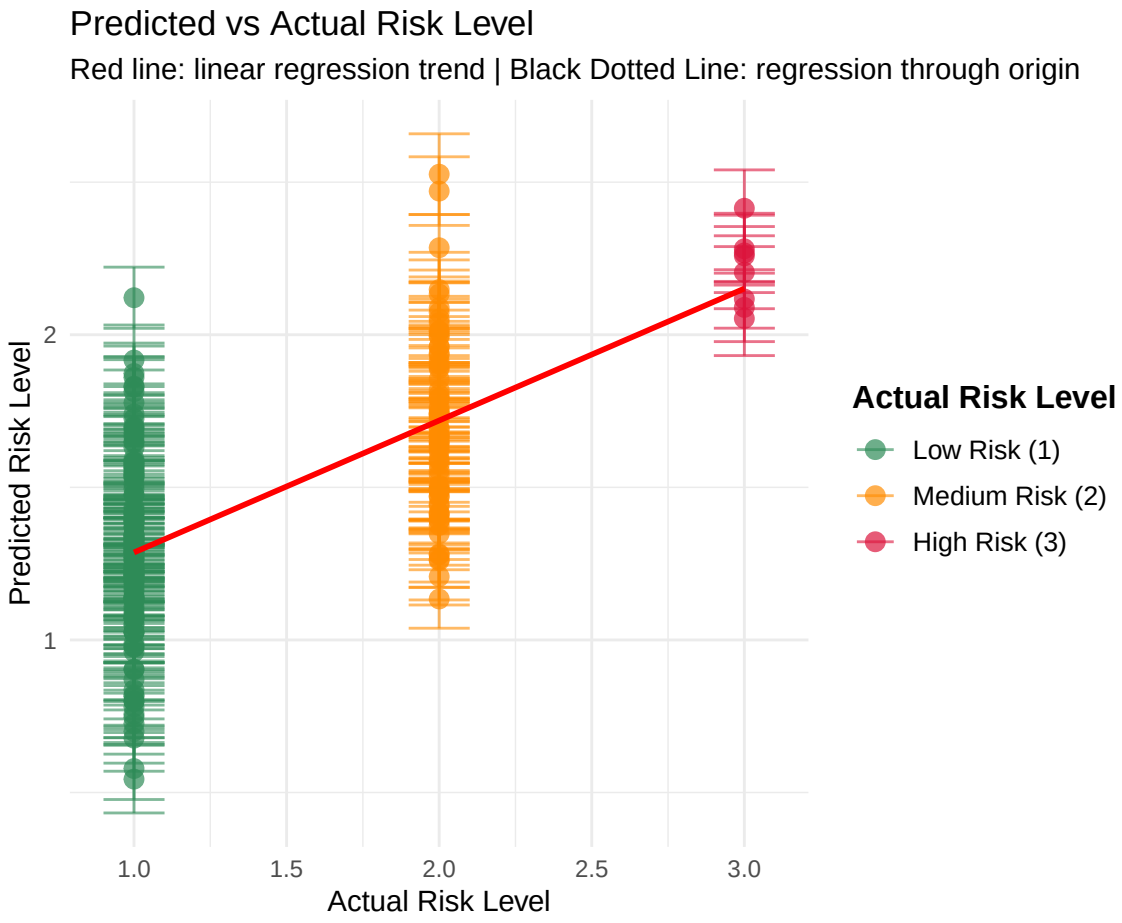
Applying the Prediction vs Actual Risk analysis to the variable smokers was not as useful because the factors were limited to yes or no so the confidence in the model's prediction is hard to tell.



As we can see when we view the regression visualized the chance is greatly different based off if you are a smoker or not. This is much easier to view than compaired to the Prediction vs Actual Risk Analysis graph.



Applying the Prediction vs Actual Risk Analysis to all the variables results in a good performance of the linear regression model in predicting against what is actually happening in the data. The points are close to the line so there is high degree of confidence in the model's predictability.


Predicted vs Actual Risk Level
Red line: linear regression trend | Black Dotted Line: regression through origin
Predicted Risk Level
2
1
1.0
1.5
2.0
2.5
3.0
Actual Risk Level
Actual Risk Level
Low Risk (1)
Medium Risk (2)
High Risk (3)