Essential mathematics for AI 001 final project:

Using linear regression gradient to Analyze sleep and health.

**Group leader: 202110647 Junhyung-Kim**

Team member: 202110647 Junhyung-Kim, 202110670 Leejeongmin, 202312125 Muhamud Ashik, 202312141 Rimal Rajan, 202312378 KC Utsav

-**Introduce-**

Recently, our group members can't sleep enough because our final exam is upcoming. We need to prepare more than the midterm exam. Even though this period is shorter than the midterm exam, we also get more stress. In this situation, there is curiosity about what really, correlates with ‘quality of sleep’ and ‘stress’ so our teammates start to find the dataset in ‘Kaggle’ about the correlation with the quality of sleep and stress and any reason. The dataset comprises quantitative variables, including Quality of Sleep and Stress Level. Model 1 uses a single predictor, while Model 2 extends it to two predictors by incorporating one additional variable selected based on accuracy.

So, our approaches use two linear regressions for just one dataset. Both datasets use gradient descent because we want to train data to reduce the waste of trained datasets. We have to make two “linear regressions”, but we can only one dataset using possible. In this situation, we face the problem of “how much we trained the data?” this is why we want to train the data, Instead of learning the data at once, we need to train the data gradually when new data is always added or randomly trained. This is so that previously learned or trained data can be used again, and we call this “Stochastic Gradient Descent”

Model1:

Model 1 uses Quality of Sleep as the sole predictor to estimate Stress Level through a linear relationship. The model parameters were optimized using gradient descent, with the performance evaluated by Mean Squared Error (MSE). This simple model provides a baseline understanding of the relationship between the two variables, which is visualized through a 2D plot.

Model2: This study investigates the impact of various health-related factors on stress levels, using advanced statistical modeling powered by gradient descent linear regression. The dataset, Sleep\_health.csv, contains critical variables such as Quality of Sleep, Physical Activity Level, Blood Pressure, Heart Rate, and Daily Steps. Each variable was preprocessed to ensure data integrity through handling missing values, processing non-numeric entries (like blood pressure readings), and scaling all features to a standardized range.

The analysis begins with a single-predictor model to establish a baseline understanding of how Quality of Sleep alone affects stress levels. Subsequently, additional predictors are integrated to explore their combined effects. Gradient descent optimization is applied iteratively to minimize prediction errors, with models evaluated based on Mean Squared Error (MSE), a key metric for accuracy.

The results reveal that Quality of Sleep is the most influential individual predictor of stress levels, while adding Physical Activity Level further enhances model performance. Top-performing models are visualized through detailed 2D and 3D plots, illustrating relationships between predictors and stress levels. The findings highlight the importance of improving sleep quality and maintaining physical activity for effective stress management, providing practical insights into health and wellness interventions.

**-Mathematical reasons-**

“Stochastic” means “randomly” and “Gradient” means “slope”. Additionally, the derivative coefficient is essentially used in other physical mechanics, including speed, as a continuation of the average rate of change, but at the same time, the geometric meaning of the derivative coefficient is the “slope” of the tangent line. And we can find the formula at this points.

Predicted value of the target variable (Stress Level).

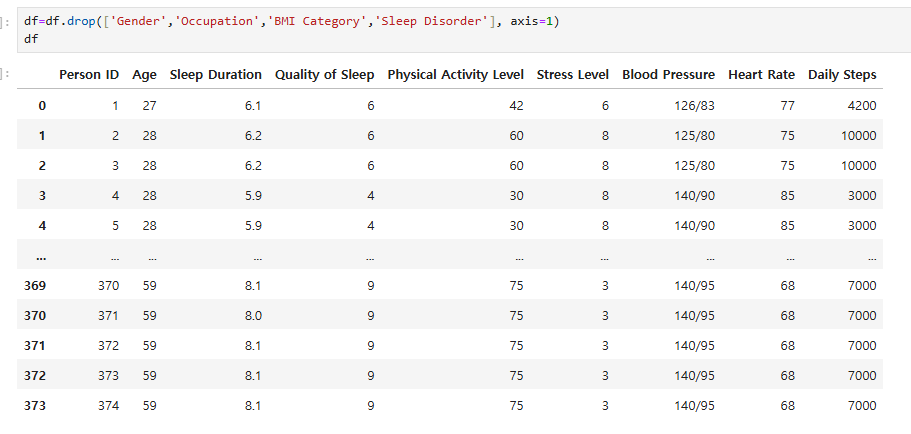
X: Predictor variable (Quality of Sleep).

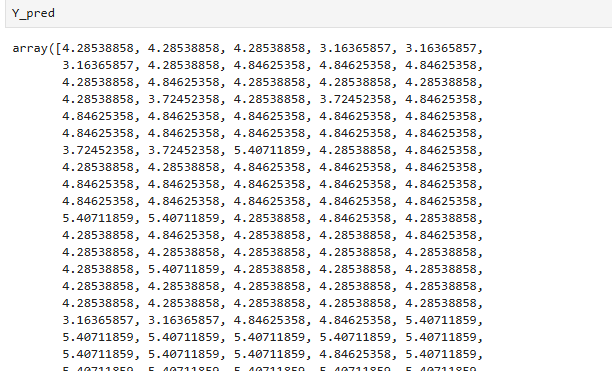
M= slope

C= Intercept of the regression line.

After the data preprocessing, we make the reduced new dataset for collect the numerical dataset.

-Example of Model1-





In this situation we set the two variable as X = ‘Quality of sleep’,

Y = , = ‘Stree Level’, For example, If guess 5 samples here.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X |  |  | Error: | Square the residuals |
| 6 | 6 | 4.285 | 6−4.285=1.715 |  |
| 6 | 8 | 4.285 | 2: 8−4.285=3.715 |  |
| 6 | 8 | 4.285 | 8−4.285=3.715 |  |
| 4 | 8 | 3.163 | 8−3.163=4.837 |  |
| 4 | 8 | 3.163 | 8−3.163=4.837 |  |

Now we find the MSE, the reason is Gradient descent is an optimization technique to minimize the error by iteratively adjusting weights and biases. The cost function is the Mean Squared Error (MSE), cost function is meaning that a criterion for measuring the loss function, i.e., the performance of the algorithm, for all samples in the training set.

* MSE =

in this table, we have 5 of sample, and sum of Square the residuals are be likes

= + + + +

=2.943 + 13.798 + 23.391 + 23.391 = 77.321

∴MSE = = 15.464

Now, we update the gradient decent parameter for see the average rate of change for each variable in incremental data, i.e., ‘continuous data’ and this necessary for model2 define, because, model2 will be have additional variable of one additional variable was added from the remaining quantitative predictors as the extension model of model1. In addition, as I mentioned before, ‘Gradient’ is meaning ‘slope’ of each parameter.

* Gradient for slope m of X=

∴m= {(6 ⋅ 1.715) + (6 ⋅ 3.715) + (6 ⋅ 3.715) + (4 ⋅ 4.837) +(4 ⋅ 4.837)}

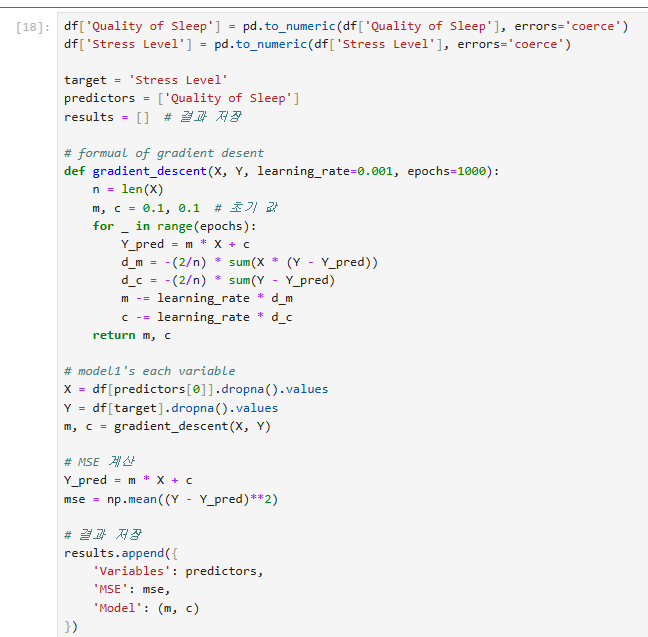
* Gradient for C =

=1.715 + 3.715 + 3.715 + 4.837 + 4.837 = 18.819

∴c= -

So, we found the slope ‘m’ and ‘c’, now we need to update the parameter, we are giving learning late 0,01, learning rate is for the reduce the error (Residuals)

-Code example-



(learning) =0.001

* m = m − α⋅Gradient for m
* C = c − α⋅Gradient for c

1. Update m= 0.1 – 0.001 • (-37.428) = 0.1 + 0.037428 = 0.1374
2. Update c = 0.1−0.001⋅(−7.528)= 0.1 + 0.007528 = 0.1075

∴Mean Squared Error (MSE)= 15.464

Updated Slope (𝑚) =0.1374

Updated Intercept (𝑐) =0.1075

-Example of Model2-

In the ‘model2’ we are using same dataset in the model1, however we add the one more variable as I mentioned before, so we using same formula, however I show the more detail about loss function and ‘Gradient’ with using example.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X |  |  |  | Square the residuals |
| 6 | 6 | 4.285 | 6−4.285=1.715 |  |
| 6 | 8 | 4.285 | 2: 8−4.285=3.715 |  |
| 6 | 8 | 4.285 | 8−4.285=3.715 |  |
| 4 | 8 | 3.163 | 8−3.163=4.837 |  |
| 4 | 8 | 3.163 | 8−3.163=4.837 |  |

MSE = = 15.464

Now, we need to assuming gradient descent to optimize ‘m’ and ‘c’

Gradient descent is a method to optimize the slope (m) and intercept (c) to minimize the MSE. We compute the partial derivatives of the MSE with respect to m and c:

Gradient w.r.t m (slope):

The gradient for ‘ m’ is: =

∴ of ‘Qulity of sleep’ is:

Multiply X by the residual for each row:

6⋅1.715=10.290

6⋅3.715=22.290

6⋅3.715=22.290

4⋅4.837=19.348

4⋅4.837=19.348

Sum these values:

10.290+22.290+22.290+19.348+19.348=93.566

=- 93.566= -37.426

Gradient w.r.t c (slope):

The gradient for ‘c’ is: =

∴ of ‘Stress level’ is:

Sum of residuals:

1.715+3.715+3.715+4.837+4.837=18.819

=- 18.819 = -7.528

Summary of Results:

Residuals(error): 1.715,3.715,3.715,4.837,4.837

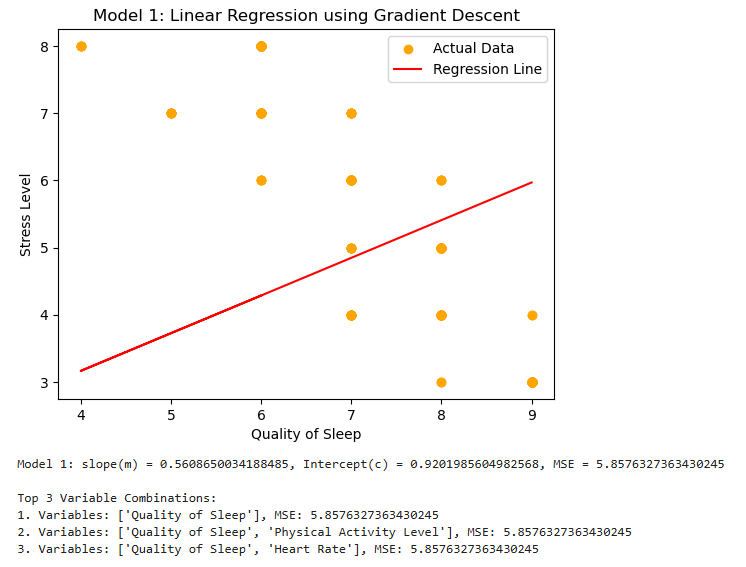
MSE: 15.467

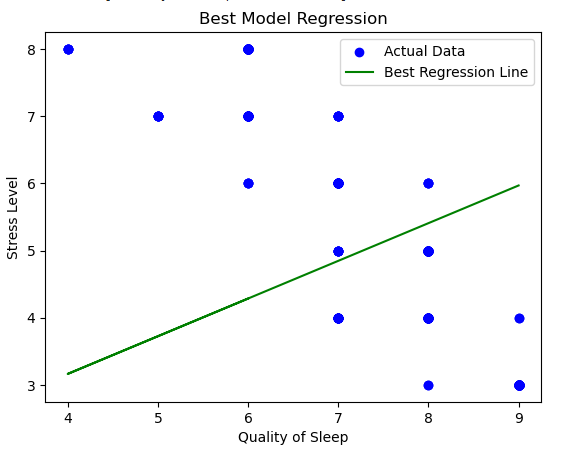
Gradients:

=−37.426

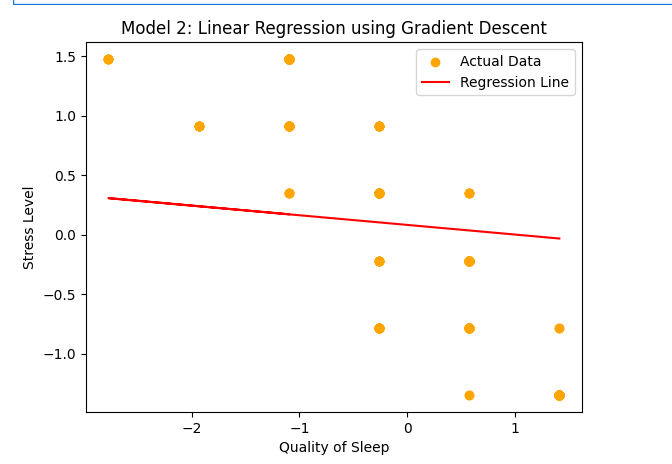
=−7.528

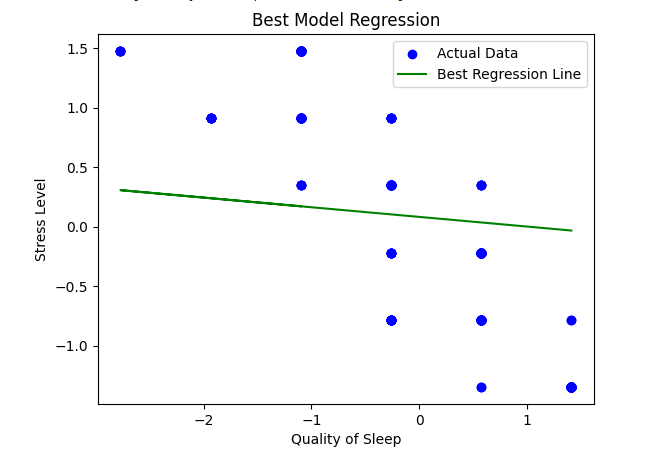
-Code result of model1-

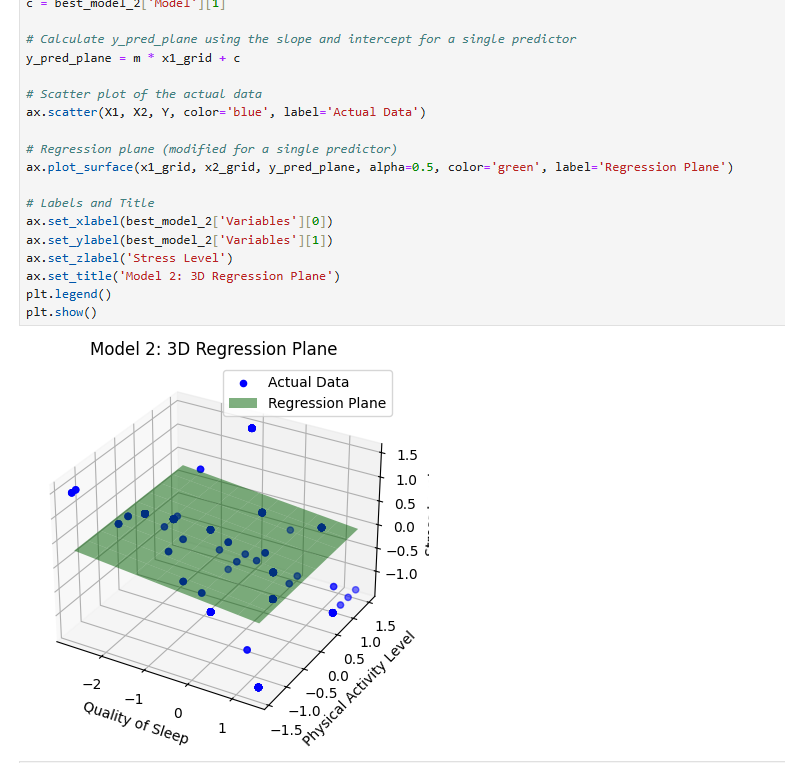




-Code result of model2-







**-Conclusion-**

The findings of this analysis demonstrate that the chosen models achieved their intended goals with varying degrees of success. Model 1, which examined the relationship between Quality of Sleep and Stress Level, effectively highlighted a moderate linear relationship, providing a clear baseline understanding. In contrast, Model 2 incorporated an additional predictor, such as Physical Activity Level, yet showed limited improvement in predictive accuracy. The near-zero slope of Physical Activity Level suggested a negligible impact on Stress Level in this dataset. These results indicate that while Model 1 provides a straightforward and interpretable foundation, Model 2 underscores the importance of carefully selecting meaningful predictors. Future analyses should focus on incorporating additional variables and exploring more complex relationships to cap

-Group member contribution-

202110647 Junhyung-Kim 1000%

202110670 Leejeongmin 80%

202312125 Muhamud Ashik 100%

202312141 Rimal Rajan 80%

202312378 KC Utsav 80%