# Diabetes Predictions Using Linear Regression Analysis

Avery Willets

University of Advancing Technology

March 31, 2025

# Diabetes Predictions Using Linear Regression Analysis

The National Center for Disease Control (CDC) estimates that 11.6% of the United States' population has diabetes (2021). With such a large part of the population having diabetes, it is critical to understand what features can indicate or lead to a progression in the disease. In this project, I explore a commonly used dataset from the paper 'Least Angle Regression' (Efron et al., 2004) that describes a collection of body and blood measurements, as well as a measurement of diabetes progression over one year.

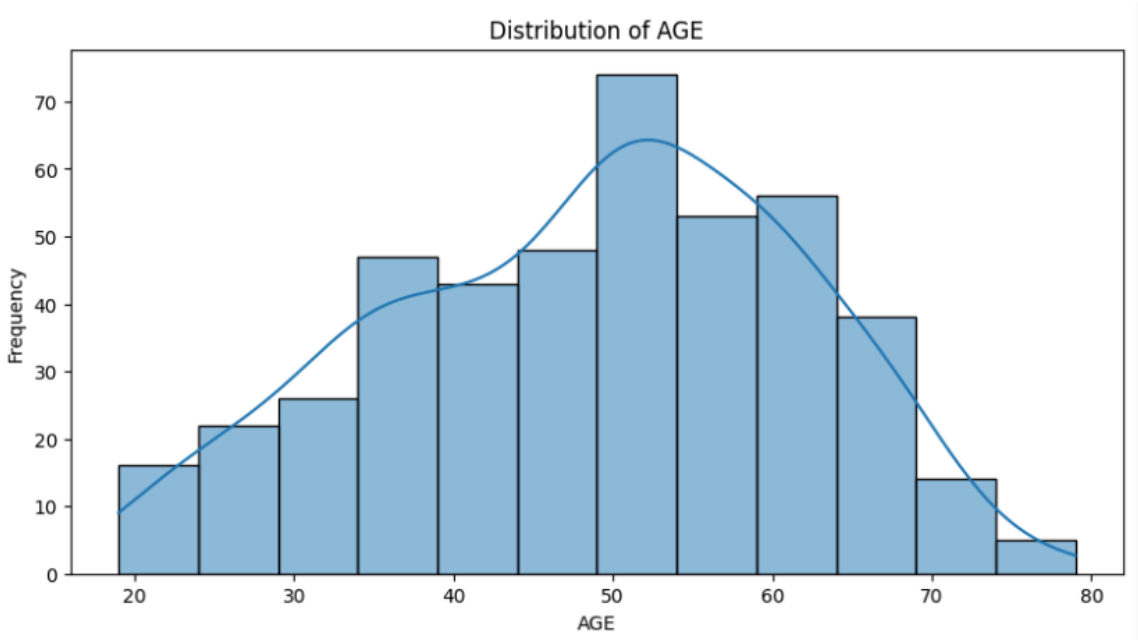
## Overview

I first explore the data to identify potential patterns within the dataset. I then implement regression analysis to predict glucose levels based on several factors. Finally, I share my results and suggest next steps for further analysis.

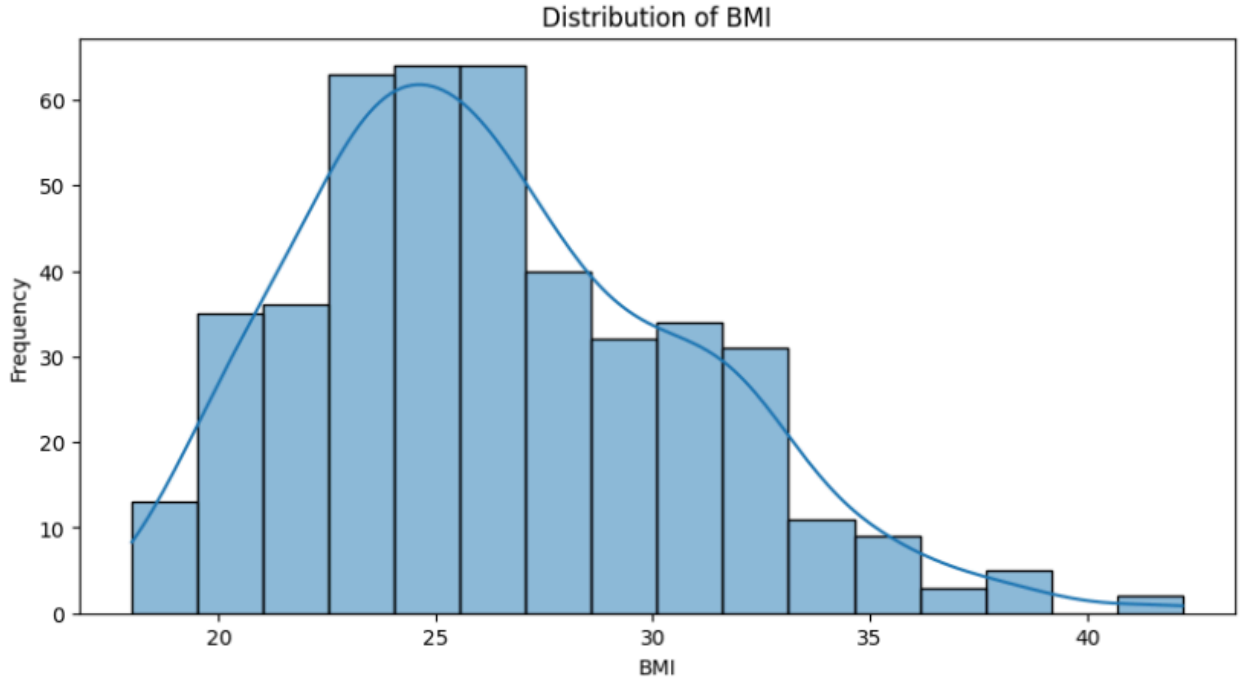
## Implementation

### The Dataset

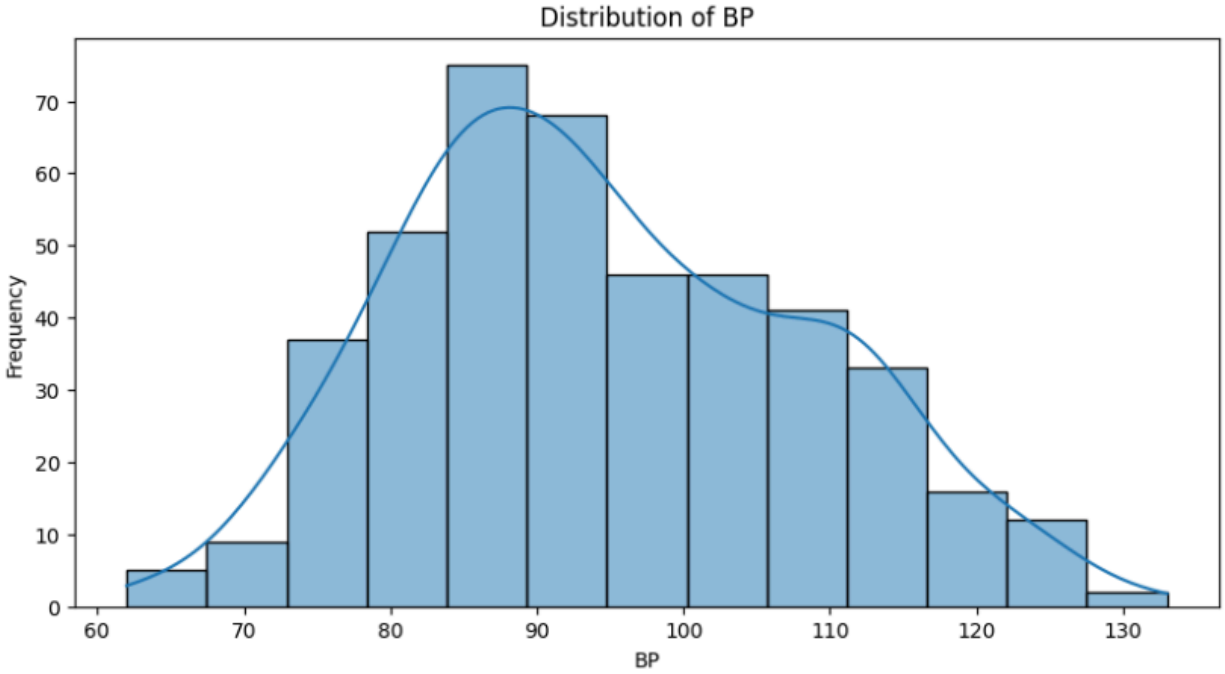
This dataset describes medical data for a sample of people with diabetes. It includes features such as age, sex, Body Mass Index, blood pressure, and six individual blood sample measurements (the specifics of which are unknown, with the sole exception being S6 measuring glucose levels), as well as a quantified measure of diabetes progression after 1 year.



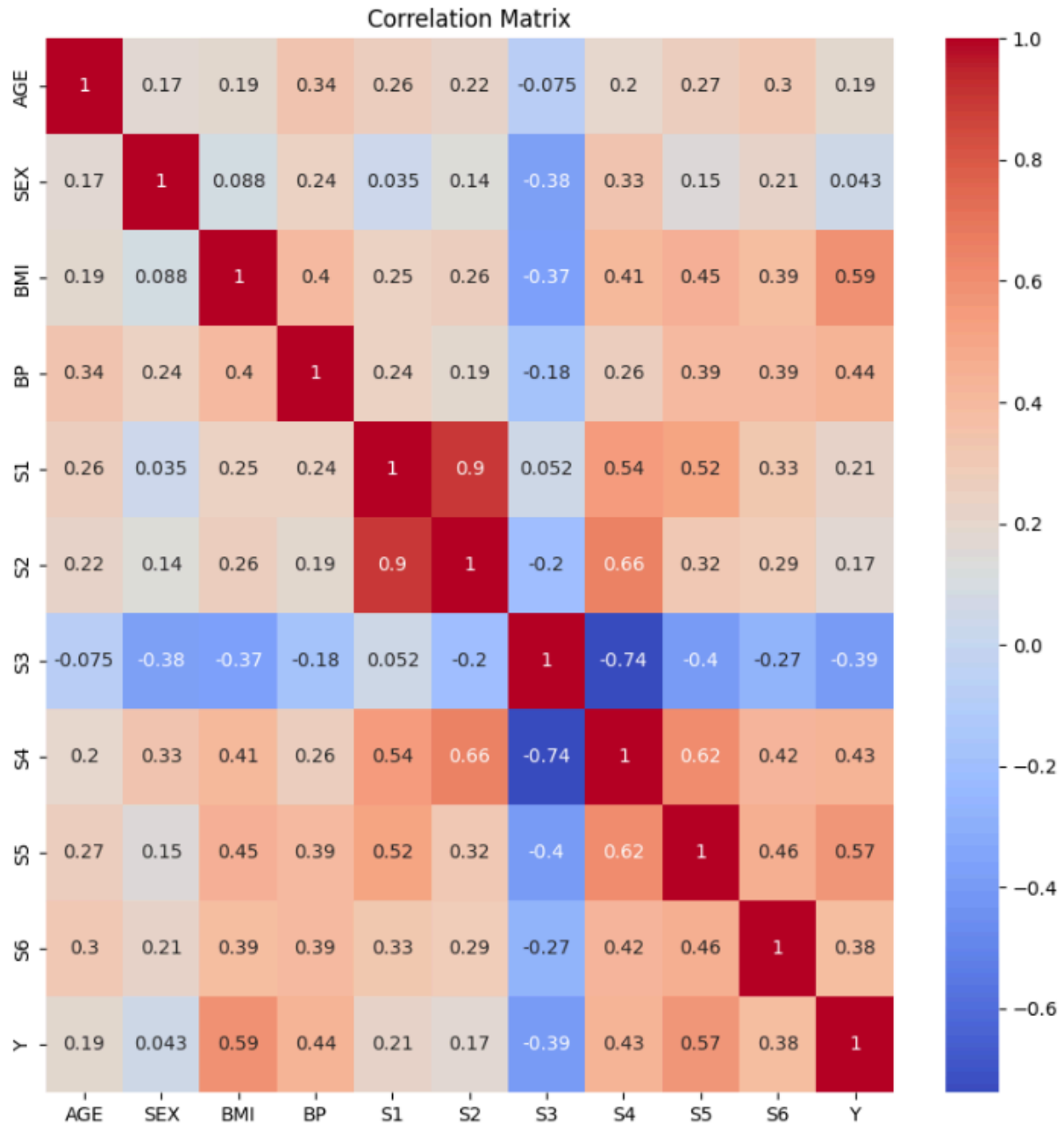

The mean age in the dataset is roughly 49, with a standard deviation of 13.1 years. This means that late 30s may be a good time to start checking thoroughly for symptoms of diabetes.




The mean body mass index in the dataset is roughly 26, with a standard deviation of 4.4.




The mean blood pressure is 94.6, with a standard deviation of 13.8. Being represented as a single integer, it likely refers to Mean Arterial Pressure (MAP). According to Nall (2021), an average MAP is anywhere from 70 to 100, so it makes sense that the mean in this dataset is within the high range of that. MAP does relate to systolic and diastolic measures of blood pressure, but I need further research on the exact calculation.




Correlating Features:

- S1 has a strong linear relationship with S2, meaning that blood sample results 1 and 2 are positively correlated.
- S2 has a relatively strong linear relationship with S4, meaning that blood sample results 2 and 4 are positively correlated.

- The features that correlated most strongly with 'Y' (The measure of progression in diabetes over 1 year) are BMI and S5. Therefore, BMI and Blood Sample 5 may be good features to explore as indicators of diabetes progression.

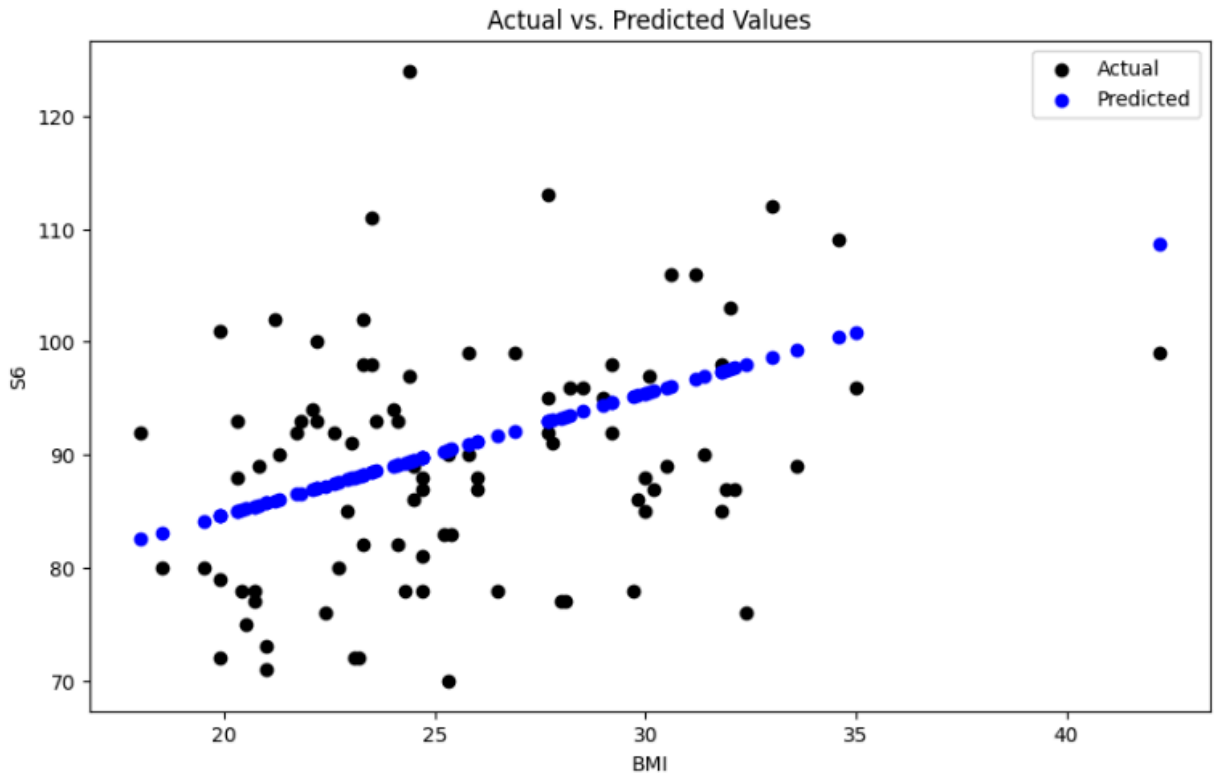
**Analysis**

***Model 1: BMI***

The first model uses exclusively BMI to predict glucose levels in blood sample 6 (S6). I train and test the model using an 80/20 split per best practices. I then fit the model to the data and visualized the predicted values. I finally evaluate the model's performance using Mean Squared Error (MSE) and R-Squared.

```
# Predicting 20% data test results
y_pred = model.predict(X_test)
y_pred
```

```
array([ 90.92220352,  95.23534704,  91.1378607 ,  96.74494727,
        89.52043188,  89.08911752, 108.60609196,  96.09797575,
        82.51157365,  93.07877528,  88.22648882,  85.74643129,
        85.42294553,  92.97094669,  85.96208847,  92.10831799,
        98.68586186, 100.41111927,  94.37271834,  97.3919188 ,
        95.45100422,  87.25603153,  86.60906   ,  93.8335754 ,
        95.99014716,  92.97094669,  97.49974739,  89.73608905,
        84.12900247,  89.41260329, 100.84243362,  87.79517447,
        92.97094669,  90.92220352,  95.55883281,  99.33283339,
        88.442146  ,  88.22648882,  88.442146  ,  85.74643129,
        84.56031683,  88.54997458,  89.3047747 ,  88.22648882,
        87.04037435,  88.01083164,  87.90300306,  86.50123141,
        83.05071659,  91.67700364,  97.3919188 ,  84.56031683,
        93.29443246,  87.4716887 ,  89.73608905,  98.03889033,
        86.93254576,  97.71540457,  89.73608905,  86.06991706,
        94.58837551,  95.66666139,  96.96060445,  87.04037435,
        84.99163118,  93.40226104,  90.38306058,  89.08911752,
        88.11866023,  90.27523199,  89.73608905,  94.58837551,
        89.41260329,  90.49088917,  85.42294553,  87.57951729,
        97.60757598,  95.45100422,  93.51008963,  89.52043188,
        91.1378607 ,  90.38306058,  85.09945977,  85.53077412,
        84.99163118,  88.98128894,  85.20728835,  84.56031683,
        95.12751845])
```




The visualization shows the BMI vs S6 glucose levels, with the predicted values represented as the blue dots and actual values as black dots. Although the line does roughly correlate with the upwards trend in the data, it fails to accurately capture the granular details.

```
Mean Squared Error: 105.68552445503605
R-squared: 0.04970798932899945
```

According to the performance metrics, the model had a very high MSE and a low r-squared value. This means the model was not very accurate in predicting S6 glucose levels based on just BMI. Therefore, at least in this dataset, BMI alone may not be the best direct indicator of diabetes.

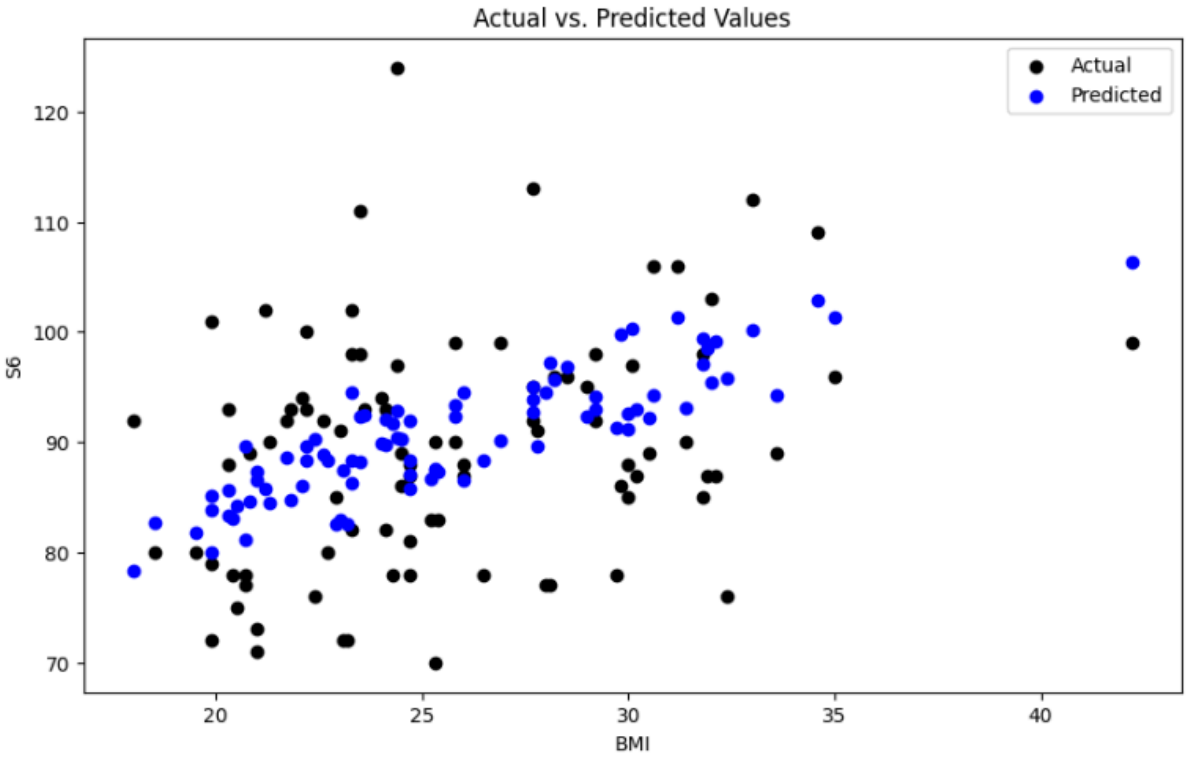
***Model 2: BMI, Age***

The second model uses BMI combined with age to predict glucose levels in blood sample 6 (S6). I train and test the model using an 80/20 split per best practices. I then fit the model to the

data and visualized the predicted values. I finally evaluate the model's performance using Mean Squared Error (MSE) and R-Squared.

```
# Predicting 20% data test results
y_pred = model.predict(X_test)
y_pred
```

```
array([ 93.37170818,  99.81461775,  94.54759703, 101.37290301,
        90.34060748,  89.75519538, 106.38288518,  94.31250526,
        78.32530953,  89.625832  ,  86.22921704,  86.54302577,
        81.14832988,  93.84632725,  85.75628618,  90.12408607,
       100.18013202, 102.91430612,  92.36338253,  99.40520992,
        91.1773644 ,  90.26020231,  84.76990732,  96.78351706,
        92.2525895 ,  95.02390432,  98.52118302,  91.90902202,
        81.74724769,  90.43958301, 101.34082694,  82.50359943,
        92.66875019,  92.39039396, 100.30274254,  94.28718205,
        88.19015727,  88.38810833,  92.311677  ,  87.32807715,
        85.08033962,  92.40896431,  91.71613561,  94.4722565 ,
        89.67310199,  87.40848232,  82.99341244,  88.59787688,
        82.73700299,  88.3610969 ,  97.05005579,  79.97750567,
        94.53071489,  88.88467419,  87.00245092,  95.86741409,
        86.04308348,  99.10828334,  88.37629083,  84.47973359,
        92.95048285,  92.94204178,  93.12817534,  88.29926208,
        83.31059759,  97.17941918,  87.58617481,  92.11034951,
        82.59919853,  86.70383612,  85.82487386,  94.12805991,
        92.79473714,  87.29093644,  89.58763218,  88.39317297,
        95.47826483,  92.55120431,  95.70660374,  90.34060748,
        86.50082042,  87.38991197,  83.01535922,  84.58208554,
        85.66575172,  89.85417091,  84.2902236 ,  83.90276255,
        91.27802814])
```




The visualization shows the BMI vs the S6 glucose levels (this time with Age being included as a factor), with the predicted values represented as the blue dots. The model's predictions were more accurate than before, indicating that age and BMI combined are better indicators of glucose levels than BMI alone. However, it was still not entirely accurate.

**Next Steps**

Further regression analysis using more features may prove beneficial. I recommend including features such as sex and blood pressure when predicting glucose levels in blood sample 6. This may allow for more accurate models with acceptable performance scores.

I also recommend using age, BMI, blood pressure, and blood sample 5 in a regression model to predict the progression of diabetes (Y). While predicting glucose levels can be beneficial in detecting diabetes, there may be other factors affecting the progression of the disease.

## References

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2). https://doi.org/10.1214/009053604000000067

Nall, R. (2021, November 30). *Understanding Mean Arterial Pressure*. Healthline. https://www.healthline.com/health/mean-arterial-pressure