Predicting Diabetes Progression Using Regression Analysis on Physiological Variables

Naga Mounica Gani - 1002199704

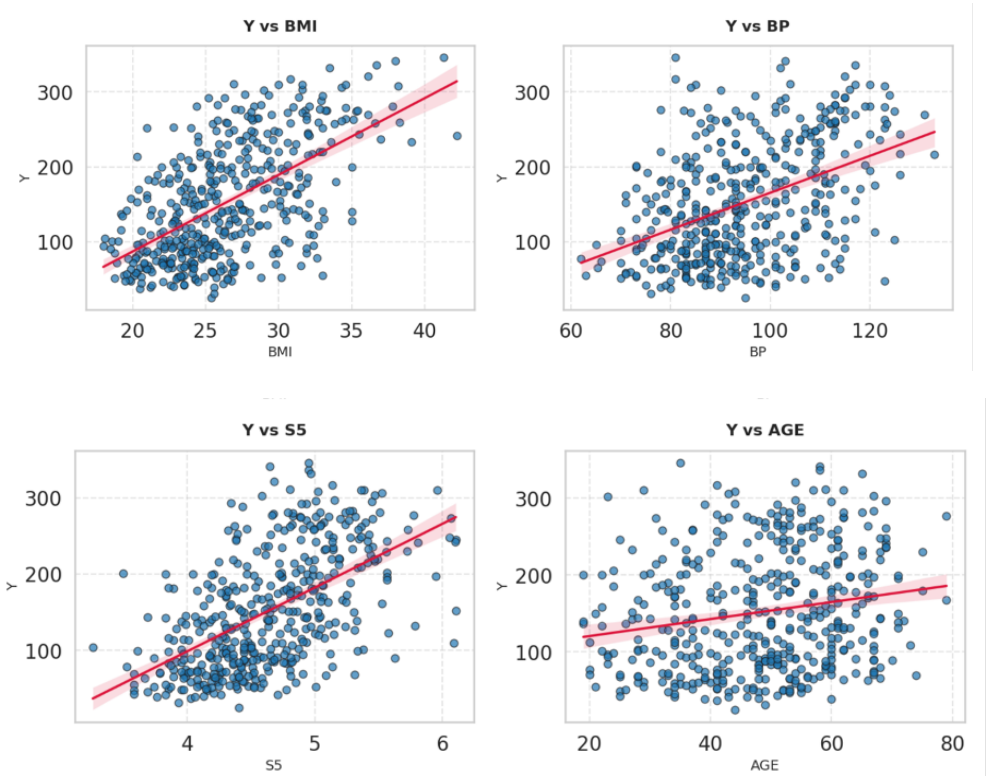
Nandini Mutha Sunil – 1002244053

1. Background and Data Description

Diabetes mellitus is one of the most prevalent chronic diseases worldwide, with hundreds of millions of people affected. According to the International Diabetes Federation, about 537 million adults (roughly 10.5% of the world's population) were living with diabetes in 2021. Most of the type 2 diabetes (T2D) patients are overweight or obese – approximately 90% of individuals with T2D have a body mass index (BMI) in the overweight or obese range. Excess body weight is a well-known risk factor for the development and progression of diabetes; in fact, being overweight or obese increases the risk of developing type 2 diabetes by 1.5 to 5 times compared to having normal BMI. These observations underscore the practical importance of understanding how BMI, a modifiable measure of adiposity, relates to the progression of diabetes. Effective weight management could potentially slow disease progression, improve glycaemic control, and reduce the risk of complications in diabetic patients.

The BMI (Body Mass Index) is a continuous variable defined as weight in kilograms divided by height in meters squared (kg/m^2). It is a widely used indicator of obesity and metabolic health. Higher BMI generally reflects higher body fat, which can exacerbate insulin resistance and inflammation, thereby potentially worsening diabetes outcomes. In this study, we examine whether BMI can serve as a useful predictor of diabetes progression over time. The response variable, denoted as Y, is a quantitative measure of diabetes disease progression one year after baseline. This progression measure (sometimes referred to as a disease progression score) is derived from clinical data and reflects the change or severity in the patient's diabetic condition after one year. A higher Y value indicates a greater progression (worsening) of the disease. While the exact scale of Y is abstract, it is a continuous index that increases with factors like elevated blood glucose, complications, or other signs of disease advancement.

For our analysis, we utilize the well-known Diabetes dataset. This dataset contains observations on 442 patients with diabetes, including 10 baseline predictor variables (age, sex, BMI, average blood pressure, and six blood serum measurements labelled S1 through S6) and the response variable Y. Each patient's data was collected at baseline, and Y was recorded one year later, making it a suitable measure of disease progression.



BMI (Body Mass Index): a numeric predictor representing the patient's body mass index (kg/m²) at baseline. This variable captures the level of adiposity; higher BMI indicates overweight or obesity status.

Y (Disease Progression): a numeric response representing the diabetes progression score one year after baseline. This score is a composite measure of diabetes progression for example, it may be based on changes in clinical metrics such as blood glucose, HbA1c, or development of complications. Higher scores mean the disease has progressed more over the year.

Other variables in the dataset include Age (years), Sex (coded 1 = male, 2 = female), BP (average blood pressure), and six serum biochemical measures (S1 through S6). These include indicators like cholesterol levels (e.g., S1 for TC - total cholesterol, S2 for LDL - low-density lipoprotein, S3 for HDL - high-density lipoprotein.), triglycerides (S5), and blood glucose level (S6). While these factors also potentially influence diabetes outcomes, in this report we focus on BMI as the sole predictor in a simple linear regression model. We chose BMI due to its practical relevance BMI is a modifiable risk factor that clinicians and patients can target through lifestyle changes or medications. By isolating BMI's effect, we aim to quantify how much on average a change in BMI impacts the progression of diabetes over one year, providing insight into the importance of weight control in diabetes management.

The dataset provides a suitable range of BMI values and corresponding progression scores for analysis. In our sample of 442 patients, baseline BMI ranges from about 18.0 (near the lower healthy weight range) to 42.2 (morbidly obese), with a mean BMI of ~26.4 (borderline overweight). The one-year diabetes progression scores (Y) range from 25 to 346 (in arbitrary units), with a mean around 152.1. This indicates substantial variability in disease progression among patients some experienced little to no progression, whereas others' conditions worsened significantly. We expect that higher BMI patients tend to have higher progression scores on average, consistent with epidemiological evidence linking obesity to worse diabetes outcomes. However, individual outcomes also depend on many other factors (genetics, diet, medications, etc.), so we anticipate variability around the trend. The following sections detail the modelling of this relationship, the statistical inferences drawn, and an assessment of the regression assumptions.

1. The Simple Linear Regression Model

The model form is linear regression. In OLS regression, the equation is:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

where y is the dependent variable. The value of y when all the independent variables are 0 is known as the intercept, or β_o. The regression coefficients for the independent variables (x_i) are β_i (i = 1, 2 ..., n). Where ε is the error term and x_i are the independent variables. To investigate the relationship between BMI and diabetes progression, the model posits a linear relationship of the form:

$$\mathbf{y = \beta_o + \beta_1 x_1 + \varepsilon}$$

where:

- y represents the quantitative measure of diabetes disease progression one year after baseline
- x_1 denotes the patient's Body Mass Index (kg/m²)
- β_o is the y-intercept parameter representing the expected progression score when BMI equals zero
- β_1 is the slope parameter representing the expected change in progression score per one-unit increase in BMI
- ε represents the random error term, assumed to follow a normal distribution with mean zero and constant variance σ^2

Parameter Estimates and Interpretation

The **fitted simple linear regression model** for the Diabetes dataset is expressed as:

$$\hat{Y} = -117.77 + 10.23 \times \text{BMI}$$

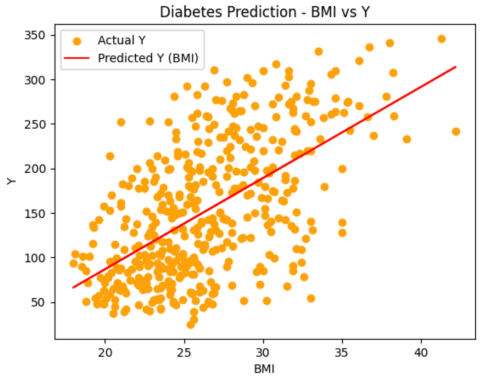
The analysis indicates a positive and statistically significant relationship between BMI and disease progression. The following parameter estimates were calculated using the OLS model of the dataset.

```
Coefficient: [10.233128]
Intercept: -117.77336656656527
```

The estimated **slope coefficient b_1 = 10.2331** suggests that, on average, for every one-unit increase in BMI, the predicted disease progression value increases by approximately 10.23 units, assuming all other factors remain constant. This positive association implies that individuals with higher BMI tend to exhibit greater disease progression, which aligns with clinical understanding that increased body mass is often linked to adverse metabolic outcomes.

The **intercept b_0 = -117.7734** represents the expected disease progression value when BMI equals zero. While a BMI of zero is not realistic in a biological context, this value serves as a theoretical baseline to define the position of the regression line on the coordinate plane. Its inclusion is essential for accurately estimating the fitted line across the observed range of BMI values.

The following is a plot of the **dependent variable Y** versus the **independent variable BMI** with a fitted regression line. A clear linear pattern is observed, indicating a positive relationship between BMI and the disease progression measure Y. As BMI increases, the corresponding values of Y also tend to rise, suggesting that individuals with higher BMI generally experience greater levels of disease progression.



2. Inferences

3.1. Inferences on parameters

From the Ordinary least square model output, the 95% confidence interval for:

- b0 (intercept): [-153.187, -82.359]
- b1 (slope): [8.909, 11.557]

It is evident that the calculated value of the intercept -117.773 lies within the confidence interval. We are 95% confident that the mean disease progression value Y will range between -153.187 and -82.359 when the BMI is zero. Although a BMI of zero is not meaningful in practice, this value serves as the model's theoretical baseline.

Similarly, the calculated value of the slope 10.233 lies within its confidence interval. We are 95% confident that the mean disease progression value Y will increase by between 8.909 and 11.557 units for every one unit increase in BMI.

This reinforces the positive relationship between BMI and disease progression, indicating that individuals with higher BMI levels are more likely to experience greater disease severity.

Overall, the narrow range of the confidence intervals suggests that the parameter estimates are precise, and the model is statistically reliable, providing strong evidence of a consistent linear association between BMI and disease progression.

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.344
Model:	OLS	Adj. R-squared:	0.342
Method:	Least Squares	F-statistic:	230.7
Date:	Sun, 26 Oct 2025	Prob (F-statistic):	3.47 ×10^-42
Time:	00:52:47	Log-Likelihood:	-2454.0
No. Observations:	442	AIC:	4912.
Df Residuals:	440	BIC:	4920.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-117.7734	18.019	-6.536	0.000	-153.187	-82.359
BMI	10.2331	0.674	15.187	0.000	8.909	11.557

Omnibus:	11.674	Durbin-Watson:	1.848
Prob(Omnibus):	0.003	Jarqu ×10^-Bera (JB):	7.310
Skew:	0.156	Prob(JB):	0.0259
Kurtosis:	2.453	Cond. No.	162.

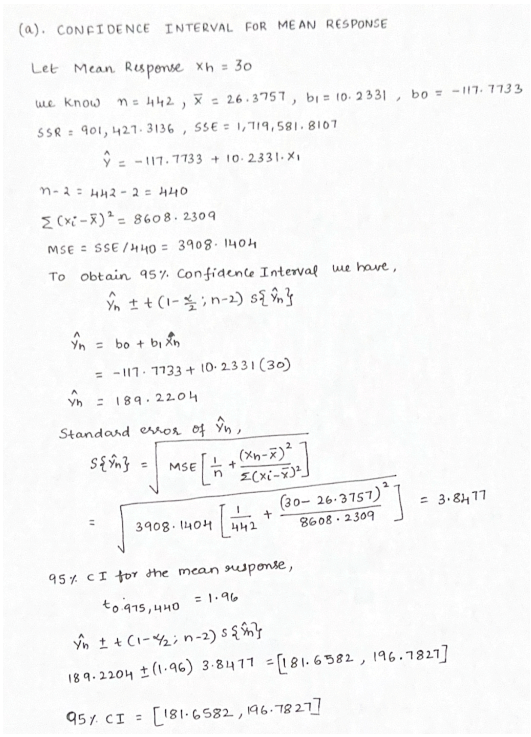
3.2. Inferences on the regression line

The following ANOVA table was used to supplement calculations of confidence intervals and prediction intervals.

ANOVA Table for BMI vs Y:

	sum_sq	df	F	PR(>F)
BMI	901427.313661	1.000000	230.653764	0.000000
Residual	1719581.810774	440.000000	NaN	NaN

Confidence Interval for the Mean Response



(a). CONFIDENCE INTERVAL FOR MEAN RESPONSE

Let Mean Response $x_h = 30$

We know $n = 442$, $\bar{x} = 26.3757$, $b_1 = 10.2331$, $b_0 = -117.7733$

$SSR = 901,427.3136$, $SSE = 1,719,581.8107$

$\hat{Y} = -117.7733 + 10.2331 \cdot X_1$

$n - 2 = 442 - 2 = 440$

$\sum (x_i - \bar{x})^2 = 8608.2309$

$MSE = SSE/440 = 3908.1404$

To obtain 95% Confidence Interval we have,

$\hat{y}_h \pm t(1-\frac{\alpha}{2}; n-2)\, s\{\hat{y}_h\}$

$\hat{y}_h = b_0 + b_1 x_h$

$= -117.7733 + 10.2331(30)$

$\hat{y}_h = 189.2204$

Standard error of $\hat{y}_h$,

$$s\{\hat{y}_h\} = \sqrt{MSE\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

$$= \sqrt{3908.1404\left[\frac{1}{442} + \frac{(30 - 26.3757)^2}{8608.2309}\right]} = 3.8477$$

95% CI for the mean response,

$t_{0.975, 440} = 1.96$

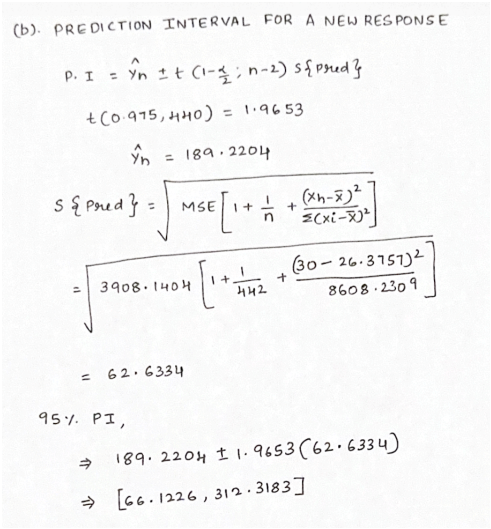
$\hat{y}_h \pm t(1-\alpha/2; n-2)\, s\{\hat{y}_h\}$

$189.2204 \pm (1.96)\ 3.8477 = [181.6582, 196.7827]$

95% CI = [181.6582, 196.7827]

The **confidence interval [181.66, 196.78]** means that we are 95% confident that the true mean disease progression Y for individuals with a BMI of 30 falls within this range. In other words, if the sampling and regression process were repeated many times and a confidence interval were computed for each sample, we would expect about 95% of those intervals to contain the true population mean response for BMI = 30.

Prediction Interval for a New Response



(b). PREDICTION INTERVAL FOR A NEW RESPONSE

$$P.I = \hat{y}_h \pm t(1-\tfrac{\alpha}{2}; n-2)\, s\{pred\}$$

$$t(0.975, 440) = 1.9653$$

$$\hat{y}_h = 189.2204$$

$$s\{pred\} = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]}$$

$$= \sqrt{3908.1404\left[1 + \frac{1}{442} + \frac{(30 - 26.3757)^2}{8608.2309}\right]}$$

$$= 62.6334$$

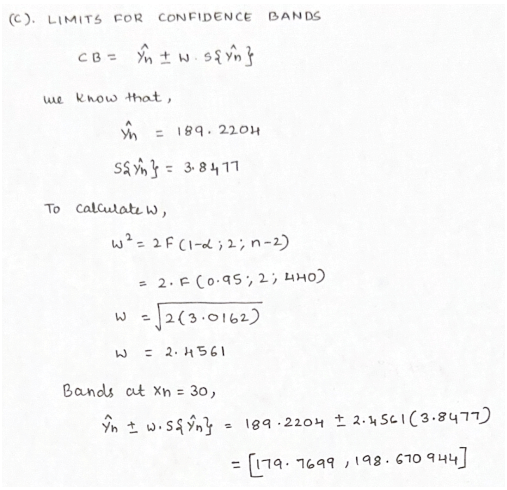
95% PI,

$$\Rightarrow 189.2204 \pm 1.9653(62.6334)$$

$$\Rightarrow [66.1226, 312.3183]$$

The **prediction interval [66.12, 312.32]** represents the range where we expect a new observation of disease progression Y to fall for an individual with a BMI of 30, with 95% confidence. Unlike the confidence interval for the mean response, the prediction interval accounts for both the natural variability in the data and the uncertainty in predicting a single new value. Therefore, it is wider than the confidence interval, reflecting the greater uncertainty associated with predicting individual outcomes rather than estimating the mean response.

Limits for Confidence Bands



(c). LIMITS FOR CONFIDENCE BANDS

$$CB = \hat{y}_h \pm W \cdot s\{\hat{y}_h\}$$

We know that,

$$\hat{y}_h = 189.2204$$

$$s\{\hat{y}_h\} = 3.8477$$

To calculate W,

$$W^2 = 2F(1-\alpha; 2; n-2)$$

$$= 2 \cdot F(0.95; 2; 440)$$

$$W = \sqrt{2(3.0162)}$$

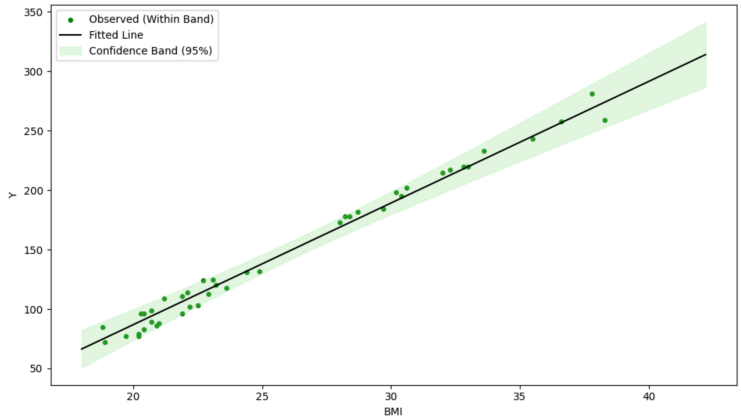
$$W = 2.4561$$

Bands at $x_h = 30$,

$$\hat{y}_h \pm W \cdot s\{\hat{y}_h\} = 189.2204 \pm 2.4561(3.8477)$$

$$= [179.7699, 198.670944]$$

The **confidence bands** provide a range where the true regression line is likely to lie at any given value of the predictor BMI. The limits **[179.77, 198.67]** indicate that we are 95% confident that the true regression line passes through this range for a BMI of 30. This means that if we were to draw multiple random samples from the population and compute confidence bands for each, we would expect about 95% of those bands to include the true regression line at xh = 30.



BMI (x_h)	Lower Limit	Upper Limit
18.0	50.755	82.09
18.834	60.502	89.422
19.669	70.215	96.788
20.503	79.884	104.198
21.338	89.497	111.664
22.172	99.034	119.206
23.007	108.471	126.847
23.841	117.777	134.621
24.676	126.911	142.564
25.51	135.835	150.72
26.345	144.513	159.12
27.179	152.933	167.779
28.014	161.105	176.686
28.848	169.063	185.806
29.683	176.848	195.1
30.517	184.498	204.529
31.352	192.046	214.06
32.186	199.517	223.667
33.021	206.931	233.333
33.855	214.299	243.043
34.69	221.633	252.787
35.524	228.94	262.559
36.359	236.226	272.352
37.193	243.494	282.163
38.028	250.748	291.987
38.862	257.991	301.824
39.697	265.224	311.669
40.531	272.448	321.524
41.366	279.666	331.385
42.2	286.878	341.252

The 95% confidence bands in the graph show that the fitted regression line captures the central trend between BMI and disease progression Y accurately. The bands are narrower near the mean BMI, indicating higher reliability of predictions where data is dense, and wider at the extremes, showing increased uncertainty. All observed points within the bands suggest the model fits those samples well.

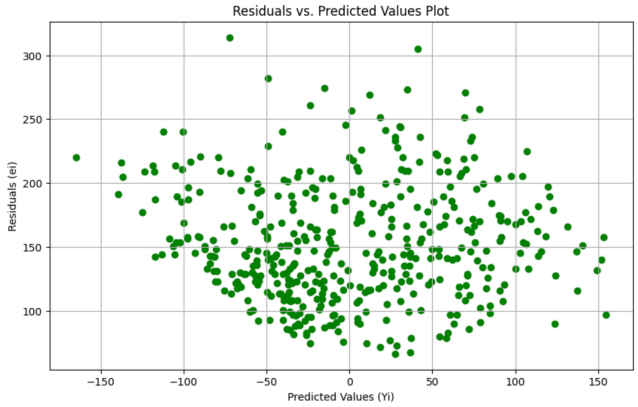
3. Model Assumptions

The below are the model assumptions

1. The linear model is reasonable.
2. The model has constant variance.
3. The normality of the model is ok.
4. There are no outliers in the model.
5. The model has uncorrelated data.

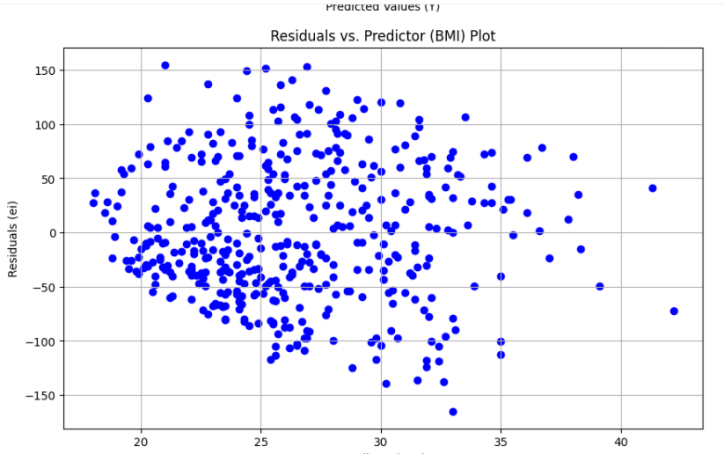
Residual Analysis Using Plots

1. Residuals vs. Predicted value (ei vs. Yi) plot



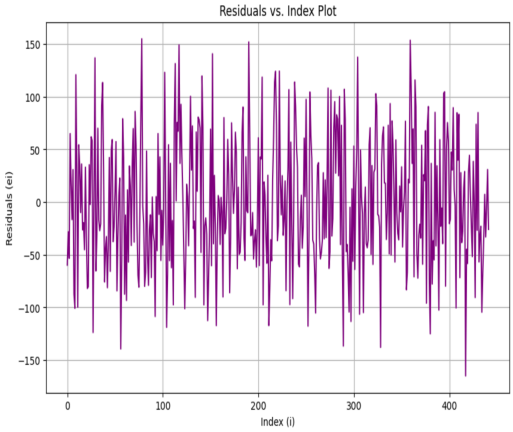
The residuals appear randomly scattered around zero without a visible pattern, indicating that the linearity and constant variance assumptions are reasonably met. However, these need to be tested for the assumption to be accurate.

2. Residuals vs. Predictor (ei vs. xi) plot



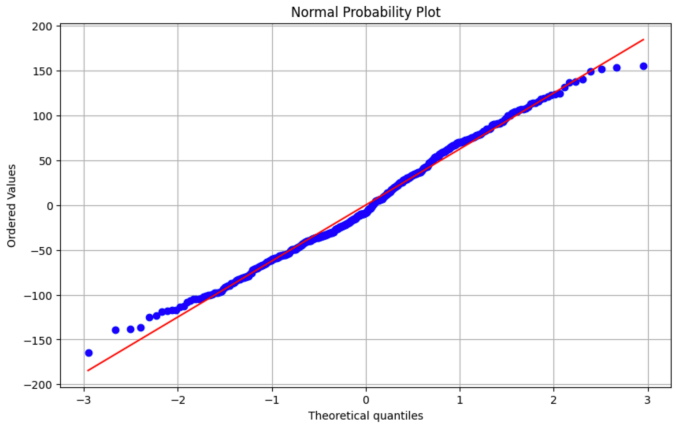
In this plot there is no clear trend, or curvature is observed and there are no strong violations of independence or model specification.

3. Residuals vs. Index (ei vs. index i) plot



The residuals are randomly distributed across the index, with no systematic pattern or clustering, suggesting that the observations are independent and errors are not correlated over time or order.

4. Normal probability plot



Most of them are approximately follow the 45-degree reference line, indicating that the residuals are approximately normal. However, a normality test needs to be conducted on the model to get a more accurate result.

Tests for normality and constant variance

Shapiro-Wilk Test (Test for normality)

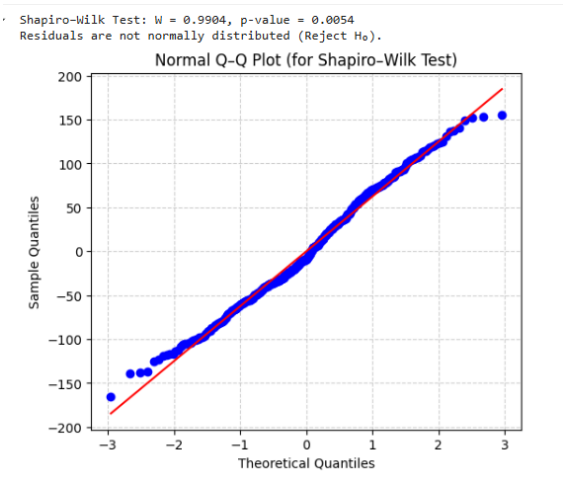
The Shapiro–Wilk test is a statistical method used to check whether a dataset follows a normal distribution. It evaluates whether the sample data come from a normally distributed population.

The null hypothesis (H_o) states that the data are normally distributed. If the calculated test statistic is less than the critical value at a chosen significance level (commonly $\alpha = 0.05$), the null hypothesis is rejected, suggesting the data are not normally distributed. However, if the test statistic exceeds the critical value, we fail to reject H_o, implying that the data may be normally distributed.

In this case,

Null hypothesis (h0): The residuals are normally distributed.

Alternative hypothesis (h1): The residuals are not normally distributed.



From the Shapiro-Wilk test, we can conclude that the data is not normally distributed.

Breusch-Pagan Test (Test for constant variance)

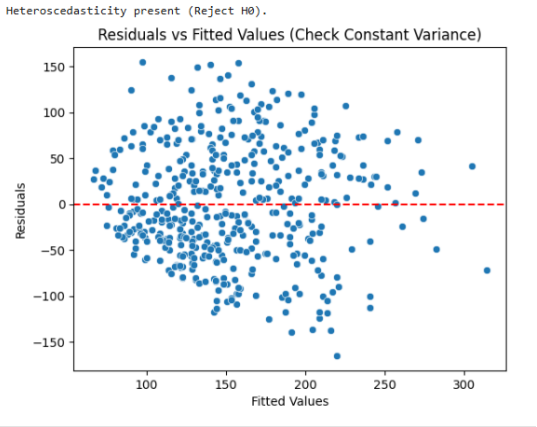
The Breusch Pagan test also known as the Breusch Pagan Godfrey test is a statistical procedure used to detect heteroscedasticity in a regression model. Heteroscedasticity occurs when the variance of the residuals (errors) is not constant across different values of the independent variables.

The null hypothesis (H_o) assumes that the residuals have constant variance (homoscedasticity). If the computed test statistic is greater than the critical value at a chosen significance level (typically $\alpha = 0.05$), the null hypothesis is rejected, indicating evidence of heteroscedasticity. However, if the test statistic is less than the critical value, we fail to reject H_o, suggesting that the residuals exhibit constant variance

In this case,

Null hypothesis (h0): The residuals have constant variance.

Alternative hypothesis (h1): The residuals do not have constant variance.



From the Breusch-Pagan test, we can conclude that the **data does not have constant variance.**

After residual analysis, the following can be concluded

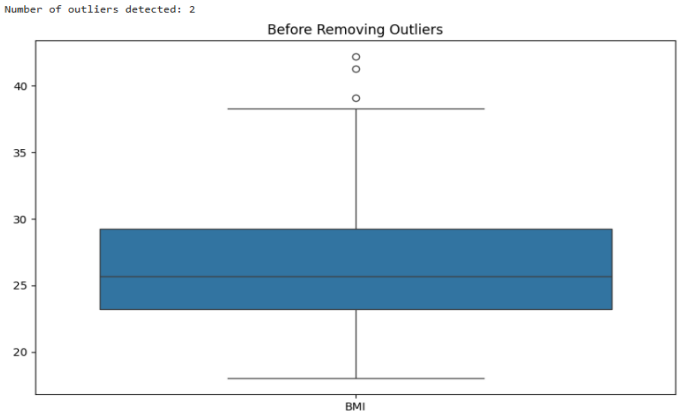
- The linear model used in this project is reasonable.
- The model does not have constant variance.
- Few outliers were found.
- The Normality of the model is not ok.
- The model has uncorrelated data.

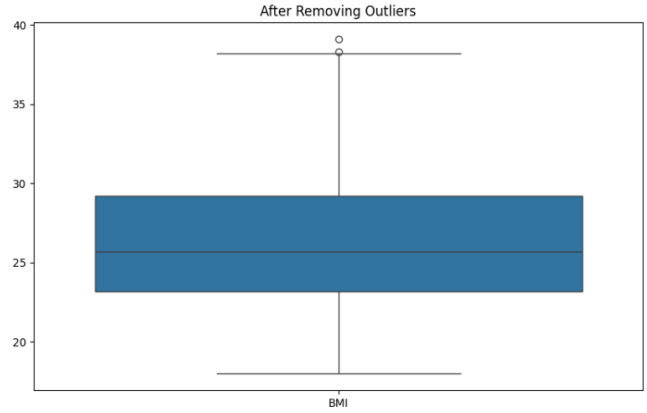
Since the model is not normal and does not have constant variance, transformation attempts need to be done.

Need for Transformations

Normality: If the data are not distributed normally, the distribution can be made more normal by using transformations such the square root, logarithmic, or Box-Cox transformations. This can guarantee the validity of the statistical conclusions made from the data and enhance the model's performance.
Constant Variance: When the variance of the errors in a regression model is not constant across all levels of the independent variables, heteroscedasticity is present. This goes against the homoscedasticity assumption, which is necessary for reliable regression coefficient estimates and sound hypothesis testing. Heteroscedasticity can be lessened, and the residuals can become more homoscedastic by using transformations (such log transformations) that stabilize the variance of the data.

Removing Outliers





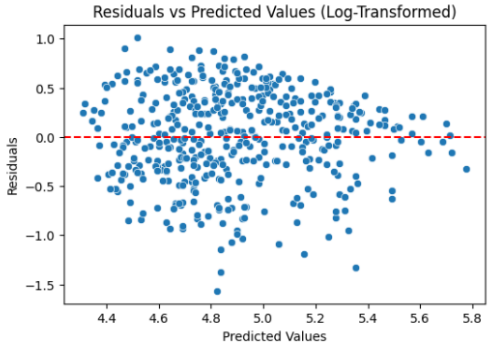

```
Data shape before removing outliers: (442, 1)
Data shape after removing outliers: (440, 1)
```

Logarithmic transformation

Logarithmic transformation is a data transformation technique commonly used to address issues related to non-normality and heteroscedasticity in statistical modelling. The primary purpose of a logarithmic transformation is to stabilize the variance of the data and normalize its distribution. Logarithmic transformation is applied to the predicted variable.

After transformation, the residual analysis is done again.

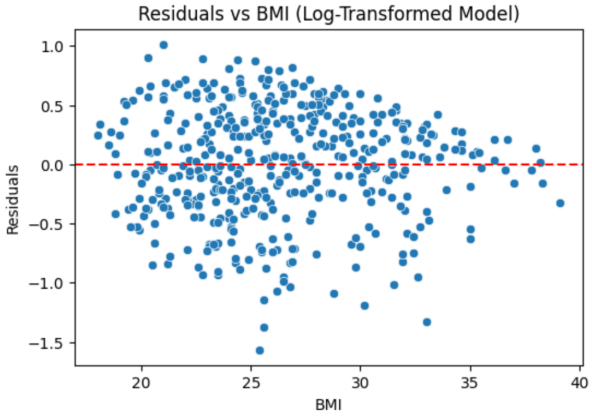
1. Residuals vs. Predicted value (ei vs. Yi) plot



The residuals are randomly scattered around the horizontal zero line without any clear pattern, indicating that the log-transformed model has improved linearity and variance stability. The spread of residuals appears consistent across BMI values, suggesting that heteroscedasticity has been reduced after transformation. Overall, the model shows a

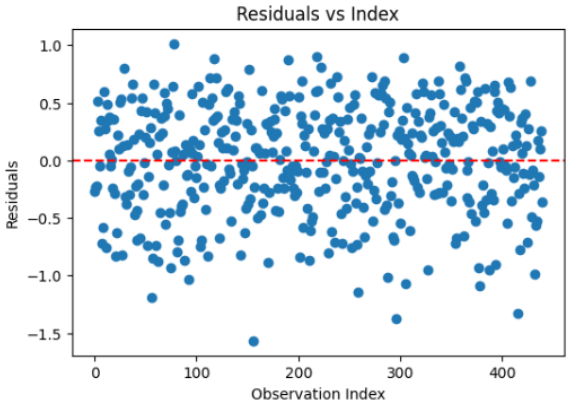
better fit with more uniform residual distribution, confirming that the log transformation effectively addressed non-linearity in the original data.

2. Residuals vs. Predictor (ei vs. xi) plot



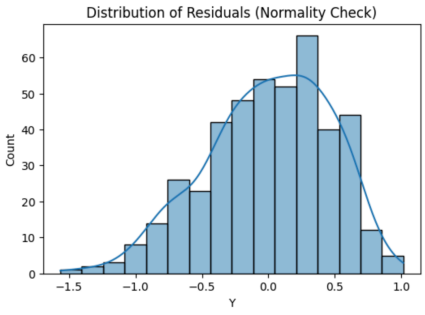
The residuals in the plot are randomly distributed around the horizontal zero line, indicating that the log transformed model provides a good linear fit between BMI and the dependent variable. There is no visible systematic pattern or curvature, suggesting that the transformation has effectively stabilised variance and improved model adequacy. Hence, the predictor BMI fits well in the transformed model.

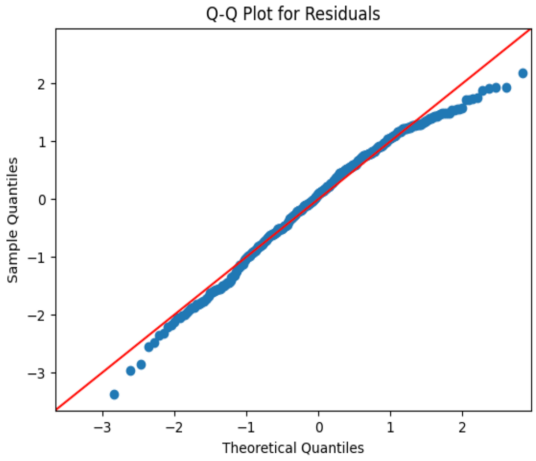
3. Residuals vs. Index (ei vs. index i) plot



The residuals are randomly scattered around the zero line without any visible trend or clustering, indicating that the model errors are independent and there is no evidence of autocorrelation. The spread of points appears consistent across the observation index, that the model's assumptions of independence and randomness of errors are satisfied.

4. Normal probability plot





The histogram of residuals displays an approximately bell-shaped distribution, indicating that the residuals follow a near-normal pattern. In the Q–Q plot, most points lie close to the 45° reference line, confirming that the normality assumption of residuals is reasonably satisfied. Therefore, the model's residuals are approximately normally distributed, supporting the validity of statistical inference in the regression model.

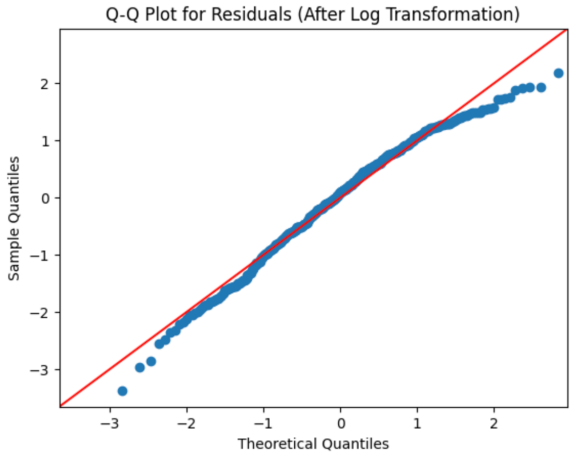
After transformation, we test for normality and constant variance again using the same tests done previously.

Based on the **Shapiro - wilk test** the normality condition of the model still does not hold good. This can be seen in the software output below.

In this case,

Null hypothesis (h0): The residuals are normally distributed.

Alternative hypothesis (h1): The residuals are not normally distributed.



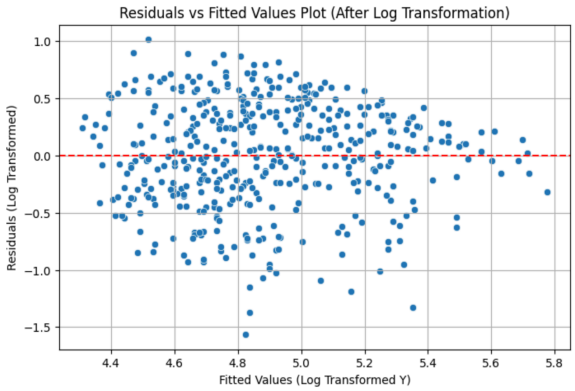

```
Original model:
  Shapiro-Wilk test statistic: 0.9903573474407812
  p-value: 0.005420877259222289
  ➤ Residuals do NOT appear to be normally distributed (reject H0).

Log-transformed model:
  R-squared: 0.9803530132282459
  Shapiro-Wilk test statistic: 0.9911806047526618
  p-value: 0.009786620718229968
  ➤ Log-model residuals do NOT appear to be normally distributed (reject H0).

Summary:
  ➤ Log-transformation improved residual normality.
```

However, upon performing the **Breusch-Pagan test,** the **model seems to have constant variance** as shown below.

In this case,

Null hypothesis (h0): The residuals have constant variance.

Alternative hypothesis (h1): The residuals do not have constant variance.




```
Breusch–Pagan Test Results after Log Transformation:
Lagrange multiplier statistic: 2.3622
p-value: 0.1243
f-value: 2.3642
f p-value: 0.1249
Constant variance assumption holds after log transformation (Fail to reject H0).
```

Summary and Conclusion

The analysis confirmed that Body Mass Index (BMI) has a strong positive relationship with diabetes disease progression. The simple linear regression model showed that as BMI increases, the progression score also increases significantly. The slope estimate indicated that every one-unit rise in BMI leads to an approximate ten-unit increase in disease progression. Initial residual analysis revealed issues with normality and constant variance, as indicated by the Shapiro Wilk and Breusch Pagan tests. To address these issues, a logarithmic transformation was applied to the dependent variable, and outliers were removed. After transformation, the residual plots displayed improved patterns with reduced heteroscedasticity and better linearity. The Breusch Pagan test confirmed that constant variance was achieved after transformation. Although the Shapiro Wilk test still indicated slight deviations from normality, the model showed overall improvement. The refined model provided reliable parameter estimates and an adequate fit to the observed data. Overall, the study demonstrated that BMI is a significant and practical predictor of diabetes progression, supported by sound statistical evidence.

Appendix

Team member roles and responsibilities:

Naga Mounica Gani:

- Working proposal and dataset
- Background & data description
- Inferences on parameters
- Simple linear regression model – interpretation and software outputs
- Hand calculations of confidence intervals, prediction intervals, and confidence bands
- Residual analysis with plots
- Drafting the report

Nandini Mutha:

- Residual analysis with plots
- Tests for normality & constant variance
- Transformations and remedial analysis
- Detection & removal of outliers
- Drafting the report

Code for finding bo and b1

```
import pandas as pd
from sklearn.linear_model import LinearRegression

df = pd.read_csv('/content/diabetes.csv')

# Select the predictor and response variables
X = df[["BMI"]]
y = df["Y"]

# Create a linear regression model
model = LinearRegression()

# Fit the model to the data
model.fit(X, y)

# Print the coefficients
print("Coefficient:", model.coef_)
print("Intercept:", model.intercept_)
```

Code for model and fitted line

```
import pandas as pd
from sklearn.linear_model import LinearRegression

df = pd.read_csv('/content/diabetes.csv')

# Select the predictor and response variables
X = df[["BMI"]]
y = df["Y"]

# Create a linear regression model
model = LinearRegression()

# Fit the model to the data
model.fit(X, y)

# Print the coefficients
print("Coefficient:", model.coef_)
print("Intercept:", model.intercept_)
```

Code for Anova table

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Fit the linear regression model (BMI as predictor, Y as response)
model = ols("Y ~ BMI", data=df).fit()

# Generate the ANOVA table
anova_table = sm.stats.anova_lm(model, typ=2)

# Disable scientific notation for clearer output
pd.set_option('display.float_format', '{:.6f}'.format)

# Display the ANOVA table
print("\nANOVA Table for BMI vs Y:")
print(anova_table)
```

Code for OLS regression Results

```
import statsmodels.api as sm
import numpy as np
# Add a constant term to the predictor variable for statsmodels
X_with_constant = sm.add_constant(X)

# Fit the OLS model
model_sm = sm.OLS(y, X_with_constant).fit()

# Print clean model summary
summary_text = model_sm.summary().as_text()

print(summary_text)
```

Code for Residual Analysis

```
# Calculate residuals and predicted values
residuals = model.resid
y_pred = model.fittedvalues

# Residuals vs. Predicted values plot (ei vs Yi)
plt.figure(figsize=(10, 6))
plt.scatter(residuals, y_pred,color='green')
plt.xlabel('Predicted Values (Yi)')
plt.ylabel('Residuals (ei)')
plt.title('Residuals vs. Predicted Values Plot')
plt.grid(True)
plt.show()

# Residuals vs. Predictor plot (ei vs xi)
plt.figure(figsize=(10, 6))
plt.scatter(X["BMI"], residuals, color='blue')
plt.xlabel('Predictor (xi)')
plt.ylabel('Residuals (ei)')
plt.title('Residuals vs. Predictor (BMI) Plot')
plt.grid(True)
plt.show()
```

```
# Residuals vs. Index plot (ei vs i)
plt.figure(figsize=(10, 6))
plt.plot(df.index, residuals, color='purple')
plt.xlabel('Index (i)')
plt.ylabel('Residuals (ei)')
plt.title('Residuals vs. Index Plot')
plt.grid(True)
plt.show()

# Normal probability plot
plt.figure(figsize=(10, 6))
stats.probplot(residuals, dist="norm", plot=plt)
plt.title('Normal Probability Plot')
plt.grid(True)
plt.show()
```

Code for Shapiro-Wilk Test

```
import matplotlib.pyplot as plt
import scipy.stats as stats

# Residuals from your fitted model
residuals = model.resid

# --- Shapiro-Wilk Test ---
from scipy.stats import shapiro
stat, p_value = shapiro(residuals)
print(f"Shapiro-Wilk Test: W = {stat:.4f}, p-value = {p_value:.4f}")

if p_value > 0.05:
    print("Residuals appear normally distributed (Fail to reject H₀).")
else:
    print("Residuals are not normally distributed (Reject H₀).")

# --- Normal Q-Q Plot ---
plt.figure(figsize=(6,5))
stats.probplot(residuals, dist="norm", plot=plt)
plt.title('Normal Q-Q Plot (for Shapiro-Wilk Test)')
plt.xlabel('Theoretical Quantiles')
plt.ylabel('Sample Quantiles')
plt.grid(True, linestyle='--', alpha=0.6)
plt.show()
```

Code for Breusch pagan test

```
# --- Breusch-Pagan Test ---
bp_test = het_breuschpagan(residuals, model.model.exog)
bp_labels = ['Lagrange multiplier', 'p-value', 'f-value', 'f p-value']
#print(dict(zip(bp_labels, bp_test)))

if bp_test[1] > 0.05:
    print("Constant variance assumption holds (Fail to reject H0).")
else:
    print("Heteroscedasticity present (Reject H0).")

# --- Plot: Residuals vs Fitted (to visualize variance) ---
plt.figure()
sns.scatterplot(x=y_pred, y=residuals) # Use y_pred instead of pred
plt.axhline(0, color='red', linestyle='--')
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.title('Residuals vs Fitted Values (Check Constant Variance)')
plt.show()
```

Code for Outliers Detection

```
# Outlier Detection using Z-score
# Compute Z-scores
z_scores = np.abs(stats.zscore(X))
threshold = 3  # typically, Z > 3 indicates outlier
outliers = (z_scores > threshold).any(axis=1)

print(f"Number of outliers detected: {outliers.sum()}")

# Visualize before removing outliers
plt.figure(figsize=(10, 6))
sns.boxplot(data=X)
plt.title('Before Removing Outliers')
plt.show()

# Remove outliers
X_clean = X[~outliers]
y_clean = y[~outliers]

# Visualize after removing outliers
plt.figure(figsize=(10, 6))
sns.boxplot(data=X_clean)
plt.title('After Removing Outliers')
plt.show()

print(f"Data shape before removing outliers: {X.shape}")
print(f"Data shape after removing outliers: {X_clean.shape}")
```

Code for Log transformation

```
# Logarithmic Transformation
# Apply log transformation to Y
y_log = np.log1p(y_clean)  # log(1 + y) to avoid log(0)

log_model = LinearRegression()
log_model.fit(X_clean, y_log)
y_log_pred = log_model.predict(X_clean)
residuals_log = y_log - y_log_pred
```

Code for Residuals after transformation

```
# (a) Residuals vs Predicted
plt.figure(figsize=(6, 4))
sns.scatterplot(x=y_log_pred, y=residuals_log)
plt.axhline(0, color='red', linestyle='--')
plt.title('Residuals vs Predicted Values (Log-Transformed)')
plt.xlabel('Predicted Values')
plt.ylabel('Residuals')
plt.show()

# (b) Residuals vs Predictor (Example: BMI)
plt.figure(figsize=(6, 4))
sns.scatterplot(x=X_clean['BMI'], y=residuals_log)
plt.axhline(0, color='red', linestyle='--')
plt.title('Residuals vs BMI (Log-Transformed Model)')
plt.xlabel('BMI')
plt.ylabel('Residuals')
plt.show()

# (c) Residuals vs Index
plt.figure(figsize=(6, 4))
plt.scatter(range(len(residuals_log)), residuals_log)
plt.axhline(0, color='red', linestyle='--')
plt.title('Residuals vs Index')
plt.xlabel('Observation Index')
plt.ylabel('Residuals')
plt.show()
```

```
# (d) Normality Check of Residuals
plt.figure(figsize=(6, 4))
sns.histplot(residuals_log, kde=True)
plt.title('Distribution of Residuals (Normality Check)')
plt.show()

sm.qqplot(residuals_log, line='45', fit=True)
plt.title('Q-Q Plot for Residuals')
plt.show()
```

Code for Shapiro test after transformation

```
from scipy.stats import shapiro

shapiro_stat_log, shapiro_p_log = shapiro(residuals_log)
print("Shapiro-Wilk Test after Log Transformation:")
print(f"Statistic = {shapiro_stat_log:.4f}, p-value = {shapiro_p_log:.4f}")

if shapiro_p_log > 0.05:
    print("Residuals appear normally distributed after log transformation (Fail to reject H₀).")
else:
    print("Residuals are not normally distributed after log transformation (Reject H₀).")
```

Code for Breusch Pagan test after transformation

```
print("\nBreusch-Pagan Test Results after Log Transformation:")
bp_labels = ['Lagrange multiplier statistic', 'p-value', 'f-value', 'f p-value']
for i in range(len(bp_labels)):
    print(f"{bp_labels[i]}: {bp_test_log[i]:.4f}")

if bp_test_log[1] > 0.05:
    print("Constant variance assumption holds after log transformation (Fail to reject H0).")
else:
    print("Heteroscedasticity present after log transformation (Reject H0).")

# Residuals vs Fitted values plot
log_model_fit = sm.OLS(y_log, X_clean_const).fit()
y_log_pred_sm = log_model_fit.fittedvalues

plt.figure(figsize=(8, 5))
sns.scatterplot(x=y_log_pred_sm, y=residuals_log)
plt.axhline(0, color='red', linestyle='--')
plt.title('Residuals vs Fitted Values Plot (After Log Transformation)')
plt.xlabel('Fitted Values (Log Transformed Y)')
plt.ylabel('Residuals (Log Transformed)')
plt.grid(True)
plt.show()
```