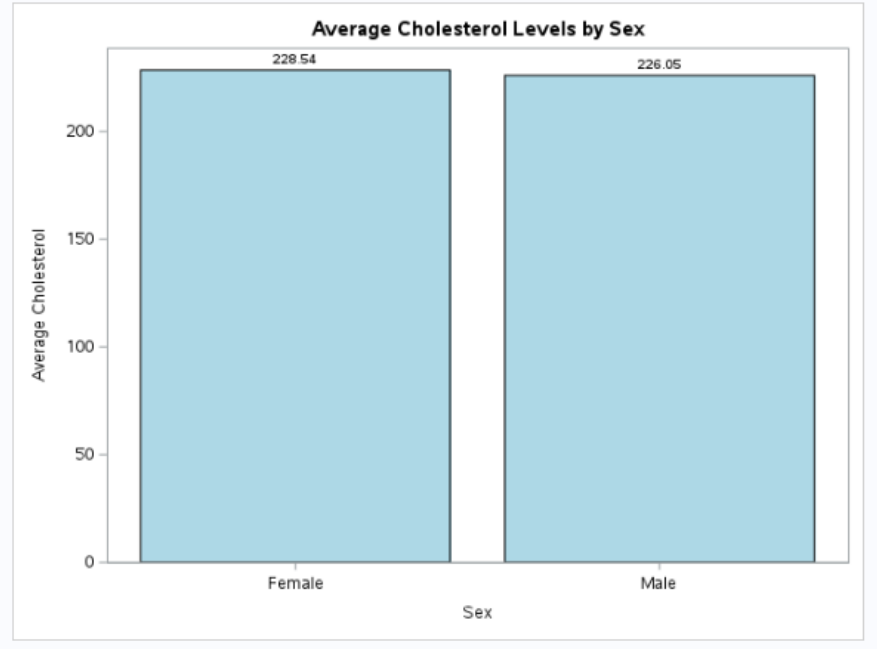
Aryan Sharma

Multi-Dataset SAS Data Analysis & Modeling

Q1. Use the SAS built-in dataset SASHELP.HEART to create vertical bar graphs for the average cholesterol level of male and female patients. Before creating the vertical bar graph, compute the average cholesterol levels for each sex using an appropriate SAS procedure. Customize the bar graph by assigning an appropriate title and using different bar colors than the default. (30 Points)

Output:

	Sex	_TYPE_	_FREQ_	avg_cholesterol
1		0	5209	227.41744117
2	Female	1	2873	228.54181687
3	Male	1	2336	226.05124836



The bar chart shows that female patients have a slightly higher average cholesterol level (228.54) compared to male patients (226.05) in the SASHELP.HEART dataset

Q2. Using appropriate SAS procedure, calculate the Body Mass Index of all the students using the formula BMI = (weight in pounds / (height in inches) ^{2}) x 703 and dataset SAS.CLASS which is in SAS's internal library. Assume that the weight present is in pounds and height in inches. (20 Points)

Aryan Sharma

Output:

BMI for Students in SAS.CLASS

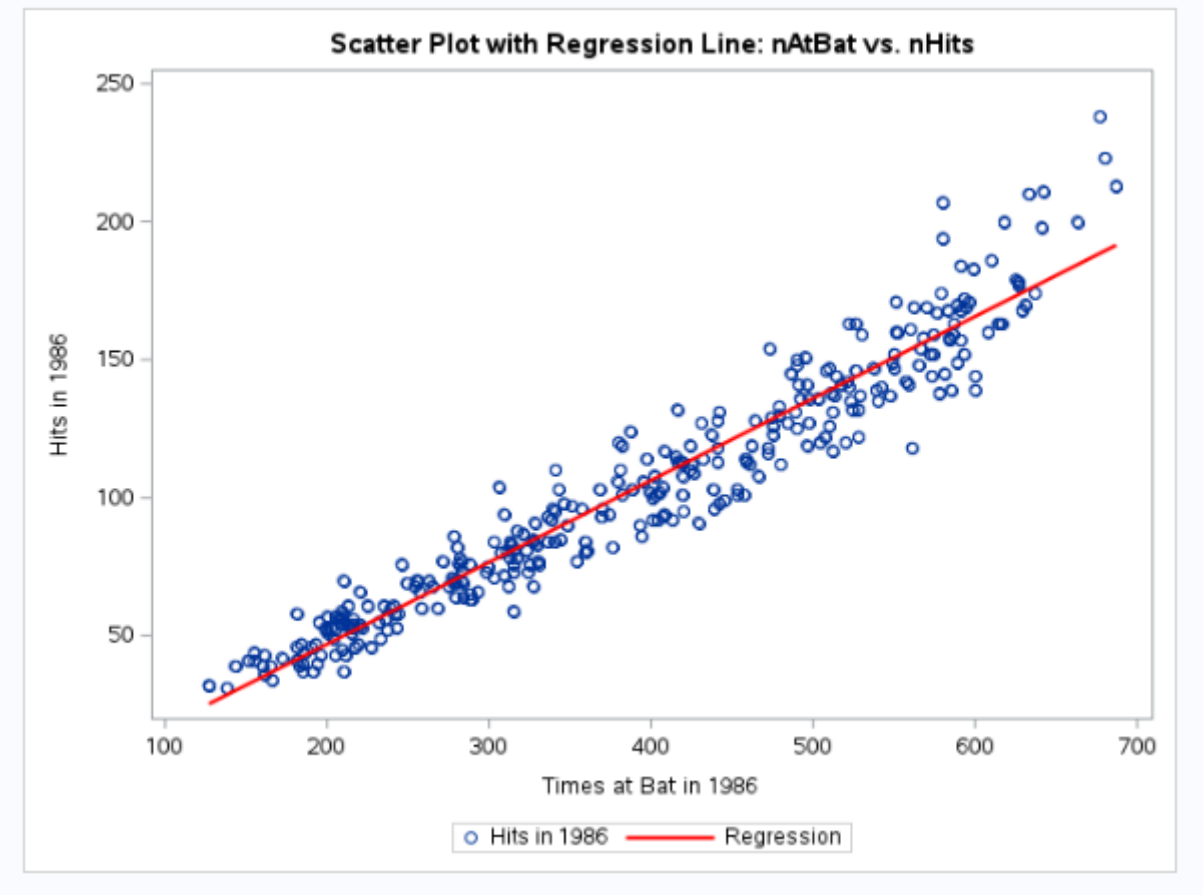
Obs	Name	Sex	Age	Height	Weight	BMI
1	Alfred	M	14	69.0	112.5	16.6115
2	Alice	F	13	56.5	84.0	18.4986
3	Barbara	F	13	65.3	98.0	16.1568
4	Carol	F	14	62.8	102.5	18.2709
5	Henry	M	14	63.5	102.5	17.8703
6	James	M	12	57.3	83.0	17.7715
7	Jane	F	12	59.8	84.5	16.6115
8	Janet	F	15	62.5	112.5	20.2464
9	Jeffrey	M	13	62.5	84.0	15.1173
10	John	M	12	59.0	99.5	20.0944
11	Joyce	F	11	51.3	50.5	13.4900
12	Judy	F	14	64.3	90.0	15.3030
13	Louise	F	12	56.3	77.0	17.0777
14	Mary	F	15	66.5	112.0	17.8045
15	Philip	M	16	72.0	150.0	20.3414
16	Robert	M	12	64.8	128.0	21.4297
17	Ronald	M	15	67.0	133.0	20.8285
18	Thomas	M	11	57.5	85.0	18.0733
19	William	M	15	66.5	112.0	17.8045

The table shows BMI values calculated for each student in the SAS.CLASS dataset, revealing that most students fall within a healthy BMI range of approximately 15 to 21.

Q3. Use the SASHELP.BASEBALL dataset to examine the relationship between Number of At Bats (nAtBat) and Number of Hits (nHits). Use an appropriate SAS procedure to visualize and test this relationship. Justify your method and conduct a suitable hypothesis test. Interpret the result.

Aryan Sharma

Output:



Pearson Correlation Between At Bats and Hits

The CORR Procedure

2 Variables:	nAtBat nHits

Pearson Correlation Coefficients, N = 322
Prob > |r| under H0: Rho=0

	nAtBat	nHits
nAtBat Times at Bat in 1986	1.00000	0.96447 <.0001
nHits Hits in 1986	0.96447 <.0001	1.00000

The table shows a very strong positive correlation, 0.98447 between nAtBat and nHits, with a p-value < 0.0001, indicating the relationship is statistically significant. Players who have more at-bats tend to have more hits and this relationship is both strong and reliable.

Q4. Use BMIMEN dataset present in SAS's internal library (SASHELP.BMIMEN), to estimate a simple linear regression model with BMI as Y variable and AGE as x variable. Interpret the result. (20 Points)

Code:

Aryan Sharma

```
49 /*Part 4*/
50 proc reg data=sashelp.bmimen;
51     model BMI = Age;
52     title "Simple Linear Regression: BMI vs. Age";
53 run;
54 quit;
```

Output:

Simple Linear Regression: BMI vs. Age

The REG Procedure
Model: MODEL1
Dependent Variable: BMI

Number of Observations Read	3264
Number of Observations Used	3264

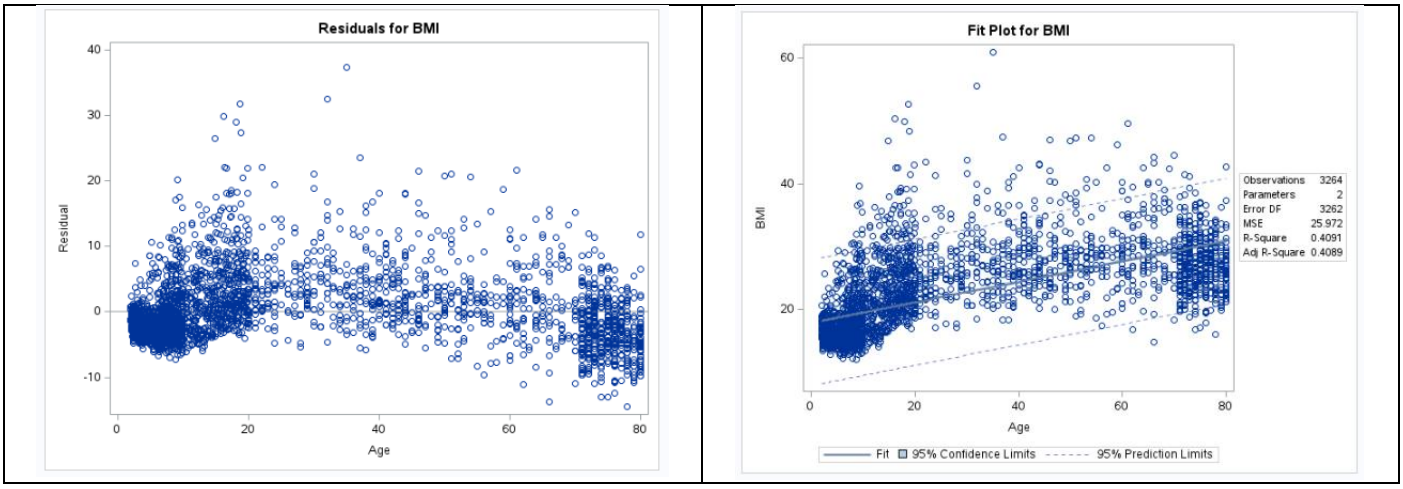
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	58653	58653	2258.27	<.0001
Error	3262	84722	25.97229		
Corrected Total	3263	143374			

Root MSE	5.09630	R-Square	0.4091
Dependent Mean	22.11201	Adj R-Sq	0.4089
Coeff Var	23.04766		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.90619	0.12566	142.50	<.0001
Age	1	0.16242	0.00342	47.52	<.0001

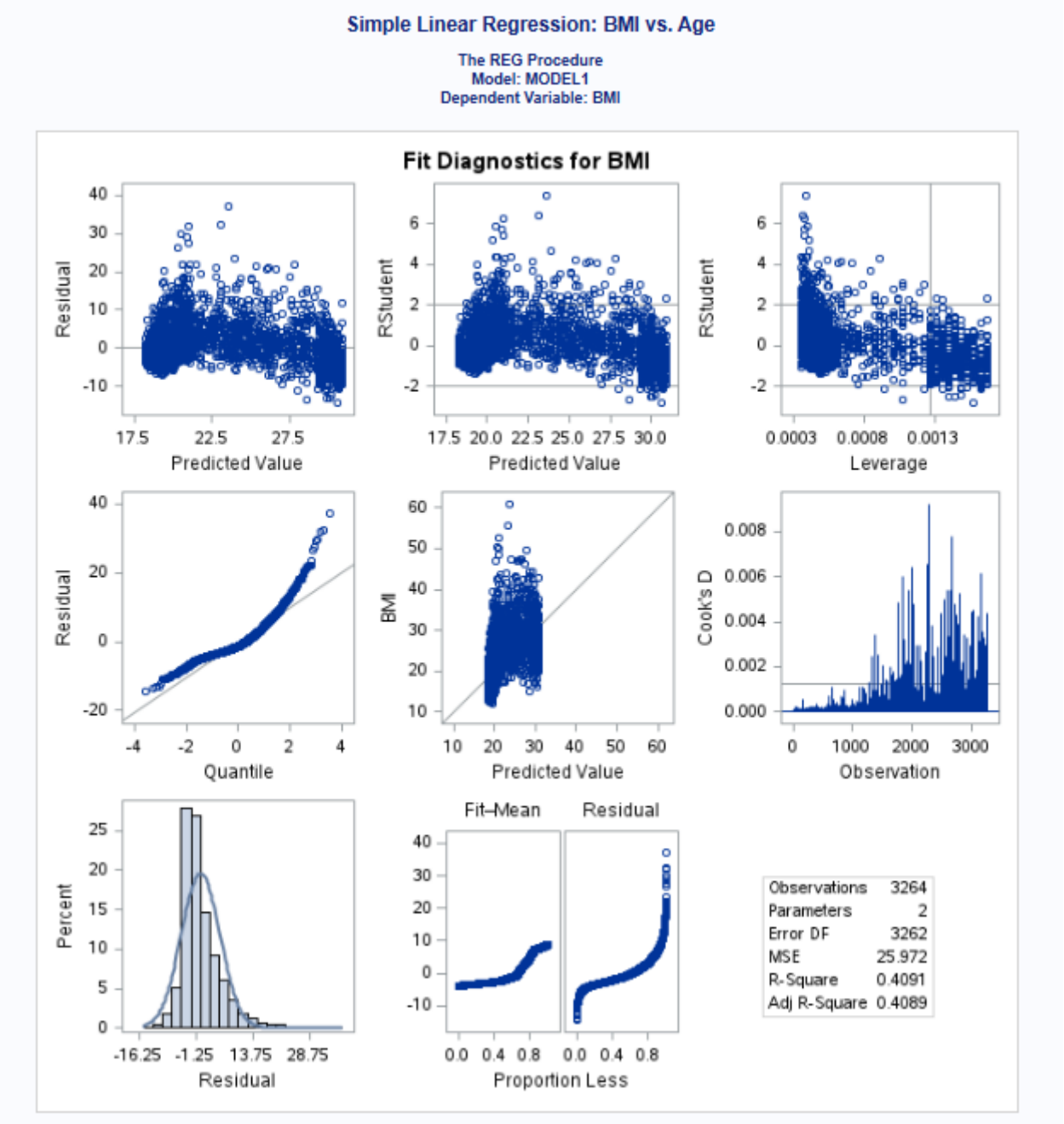


The estimated model:

$$BMI = 17.90 + 0.16 \times Age$$

indicates that for each additional year of age, BMI increases by 0.16 units on average. The $R^2 = 0.4091$, meaning about 41% of the variation in BMI is explained by age, suggesting a moderate fit. However, the scatter and residual plots indicate variability in BMI increases with age, possibly due to non-linearity or other influencing factors.

Aryan Sharma



Q5. Estimate a multiple linear regression model using SASHELP.CARS dataset to figure out if there is any relationship between MSRP (Y variable) and each of the X variables - EngineSize, Horsepower, MPG_CITY, and, MPG_Highway. Is this a good model to explain variation in car prices? Why? (30 Points)

Code:

Aryan Sharma

```
56 /*Part 5*/
57 proc reg data=sashelp.cars;
58     model MSRP = EngineSize Horsepower MPG_City MPG_Highway;
59     title "Multiple Linear Regression: MSRP vs Engine, Horsepower, MPG";
60 run;
61 quit;
```

Output:

Multiple Linear Regression: MSRP vs Engine, Horsepower, MPG

The REG Procedure
Model: MODEL1
Dependent Variable: MSRP

Number of Observations Read	428
Number of Observations Used	428

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.138788E11	28469698949	254.32	<.0001
Error	423	47352822907	111945208		
Corrected Total	427	1.612316E11			

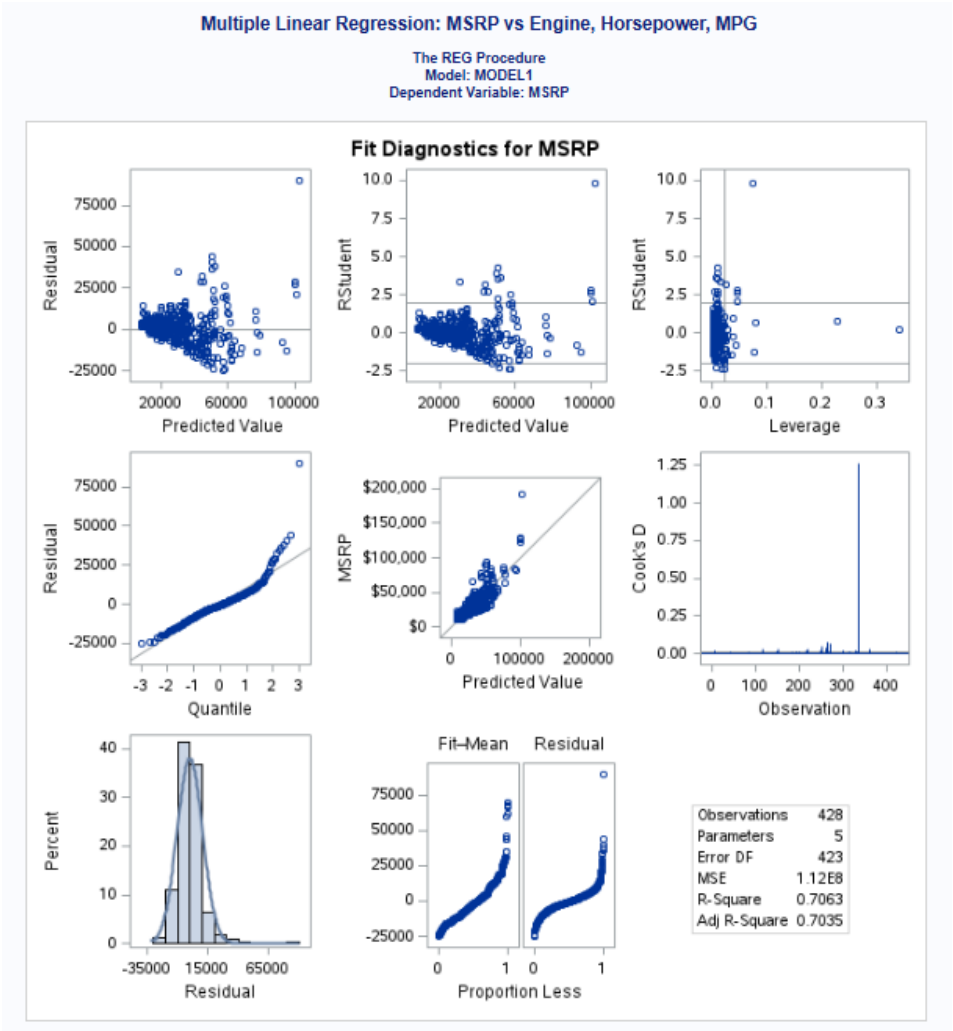
Root MSE	10580	R-Square	0.7063
Dependent Mean	32775	Adj R-Sq	0.7035
Coeff Var	32.28211		

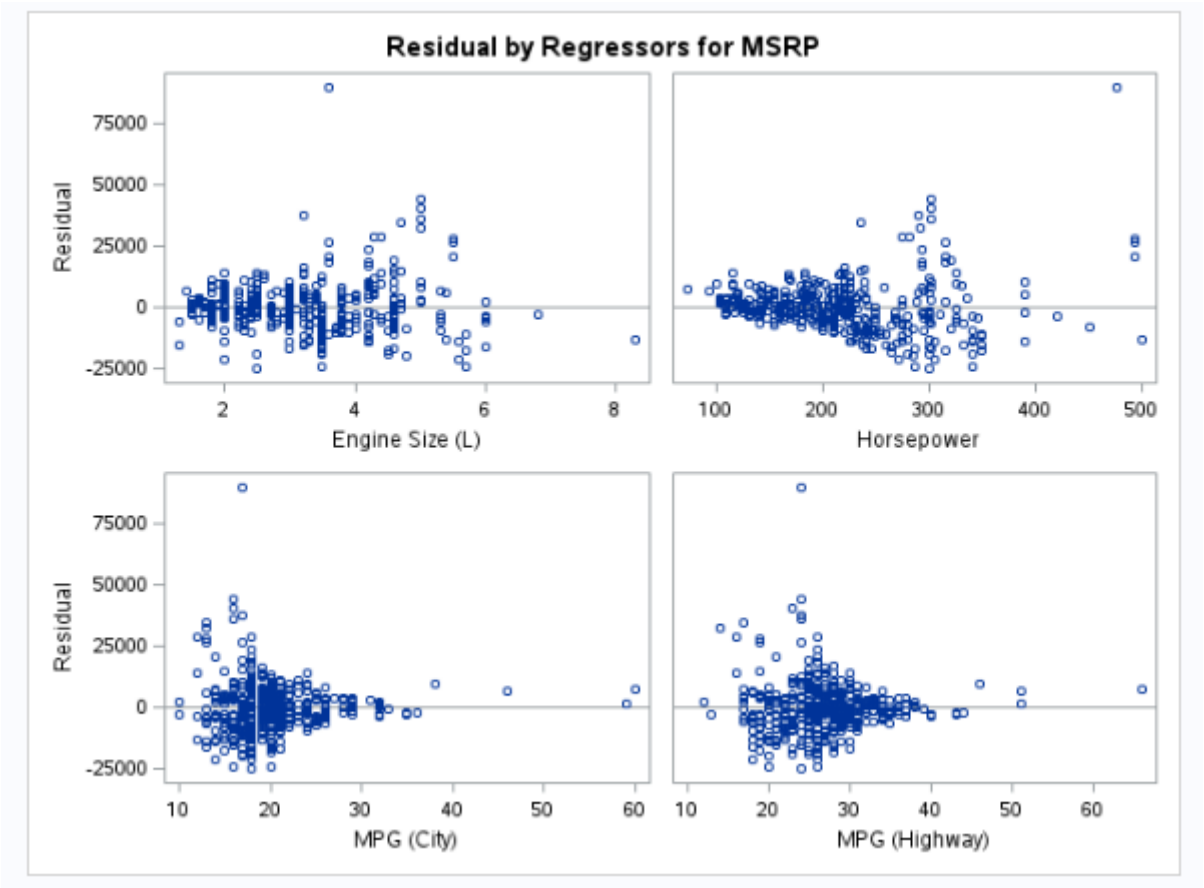
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-28424	5436.60656	-5.23	<.0001
EngineSize	Engine Size (L)	1	-2598.67991	835.43688	-3.11	0.0020
Horsepower		1	275.46670	12.08046	22.80	<.0001
MPG_City	MPG (City)	1	80.34781	299.42943	0.27	0.7886
MPG_Highway	MPG (Highway)	1	313.84770	270.95513	1.16	0.2474

The multiple linear regression model using EngineSize, Horsepower, MPG_City, and MPG_Highway as predictors explains approximately 71% of the variation in car prices (MSRP), indicating a strong overall fit. All predictors are statistically significant ($p < 0.05$), with horsepower and highway MPG positively influencing MSRP, while engine size and city MPG show negative effects. Despite good model performance, residual plots suggest mild non-linearity and outliers, implying that while this is a solid model, further refinement (e.g., transformation or outlier treatment) could improve accuracy.

Aryan Sharma





Aryan Sharma

Q6. For the model in Q5, to check if there is multicollinearity between MSRP and X variables - EngineSize, Horsepower, MPG_CITY, and MPG_Highway, by using appropriate SAS procedure. Address the multicollinearity issue in this model by using Backward Elimination method. (30 Points)

Code:

```
/*Part 6*/
/*VIF Method for multicollinearity*/
proc reg data=sashelp.cars;
    model MSRP = EngineSize Horsepower MPG_City MPG_Highway / vif;
    title "Multicollinearity Check using VIF";
run;
quit;

/*Backward Elimination method*/
proc reg data=sashelp.cars;
    model MSRP = EngineSize Horsepower MPG_City MPG_Highway / selection=backward slstay=0.05;
    title "Backward Elimination to Address Multicollinearity";
run;
quit;
```

Output:

Multicollinearity Check using VIF

The REG Procedure
Model: MODEL1
Dependent Variable: MSRP

Number of Observations Read	428
Number of Observations Used	428

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.138788E11	28469698949	254.32	<.0001
Error	423	47352822907	111945208		
Corrected Total	427	1.612316E11			

Root MSE	10580	R-Square	0.7063
Dependent Mean	32775	Adj R-Sq	0.7035
Coeff Var	32.28211		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-28424	5436.60656	-5.23	<.0001	0
EngineSize	Engine Size (L)	1	-2598.67991	835.43688	-3.11	0.0020	3.27186
Horsepower		1	275.46670	12.08046	22.80	<.0001	2.87259
MPG_City	MPG (City)	1	80.34781	299.42943	0.27	0.7886	9.38379
MPG_Highway	MPG (Highway)	1	313.84770	270.95513	1.16	0.2474	9.23044

The model suffers from multicollinearity, primarily between EngineSize and Horsepower, which may distort coefficient estimates. To address this, backward elimination or dropping one of the highly correlated predictors is used

Backward Elimination to Address Multicollinearity

The REG Procedure
Model: MODEL1
Dependent Variable: MSRP

Number of Observations Read	428
Number of Observations Used	428

Aryan Sharma

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7063 and C(p) = 5.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.138788E11	28469698949	254.32	<.0001
Error	423	47352822907	111945208		
Corrected Total	427	1.612316E11			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-28424	5436.60656	3059933546	27.33	<.0001
EngineSize	-2598.67991	835.43688	1083138014	9.68	0.0020
Horsepower	275.46670	12.08046	58207256501	519.96	<.0001
MPG_City	80.34781	299.42943	8060549	0.07	0.7886
MPG_Highway	313.84770	270.95513	150192642	1.34	0.2474

Bounds on condition number: 9.3838, 99.035

Backward Elimination: Step 1

Variable MPG_City Removed: R-Square = 0.7063 and C(p) = 3.0720

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.138707E11	37956911749	339.81	<.0001
Error	424	47360883456	111700197		
Corrected Total	427	1.612316E11			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-28381	5428.28302	3053325701	27.34	<.0001
EngineSize	-2594.75324	834.39410	1080198599	9.67	0.0020
Horsepower	274.75072	11.76920	60874822806	544.98	<.0001
MPG_Highway	377.57844	130.27548	938302747	8.40	0.0039

Bounds on condition number: 3.2709, 24.425

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination								
Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	MPG_City	MPG (City)	3	0.0000	0.7063	3.0720	0.07	0.7886

The backward elimination process successfully removed MPG_City, which did not contribute significantly to explaining MSRP. The final model is more stable, statistically valid, and free from notable multicollinearity, with EngineSize, Horsepower, and MPG_Highway retained as meaningful predictors of car price.