In order to explain the link between the price of cars (y), the peak rpm which is the biggest speed cars can reach (x_1), the horsepower which is the highest power produced by the engine (x_2), the wheelbase which is the distance between the front and back wheels' centres (x_3), the height (x_4), the width (x_5), the fuel type (x_6) and the aspiration which is the fuel consumption type of engine (x_7) of cars, the following model is fitted to the automobile data.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 \, \varepsilon \, , \varepsilon \sim \text{NIID} \, (0, \sigma^2)$$

The fitted regression equations which are different for each level of the categorical variables are as follows:

Fuel Type = Diesel, Aspiration= std

Price (y) = -69049 - 0,330 peak-rpm (x1) + 153,4 horsepower (x2) + 115,0 wheel base (x3) + 87 height (x4) + 876 width (x5)

Fuel Type = Diesel, Aspiration= turbo

Price (y)= -72933 -0,330 peak-rpm (x1) + 153,4 horsepower (x2) + 115,0 wheel-base (x3) + 87 height (x4) + 876 width (x5)

Fuel type = Gas, Aspiration = std

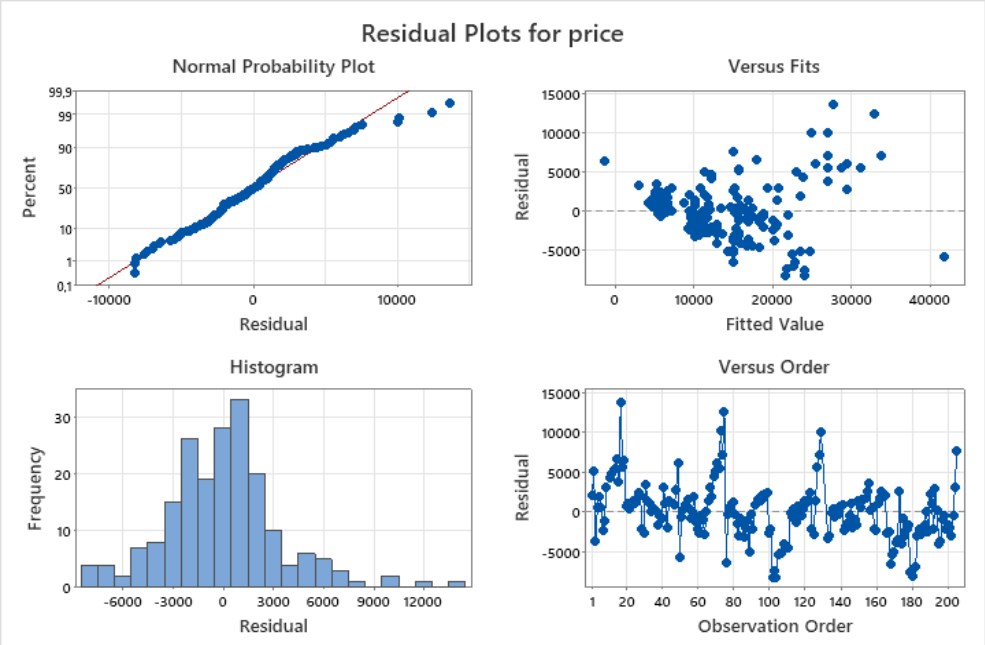
Price (y)= -74522 -0,330 peak-rpm (x1) + 153,4 horsepower (x2) + 115,0 wheel-base (x3) + 87 height (x4) + 876 width (x5)

Fuel Type = Gas, Aspiration = turbo

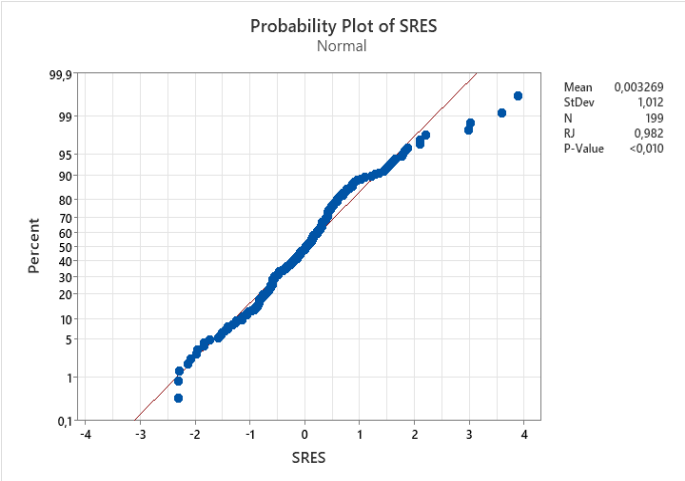
Price (y)= -78407 -0,330 peak-rpm (x1) + 153,4 horsepower (x2) + 115,0 wheel-base (x3) + 87 height (x4) + 876 width (x5)

(APPENDIX 1 MINITAB output)

To feel confident about the validity of the model, assumptions are checked by using the following residual analysis and more:



1. **Normality of Errors**
   According to the Normal Probability Plot of Residuals above, it is seen that residuals lie on the reference line closely. However, there seem to be some outlier observations, whose residuals are greater than 10000.



To be confident about whether there is a normality problem with the residuals, Ryan-Joiner (RJ) test in MINITAB is used. A decision is made according to the acceptance of the null and alternative hypotheses below:

H_0: Errors are distributed normally
H_1: Errors are not distributed normally

Since the p-value associated with the RJ test given in the NPP of residuals is smaller than 0.005 and the default Type-I probability $\alpha = 0.05$, we do reject the null hypothesis. As a result, we can say that residuals do not follow a normal distribution.

2. **Constant Variance**

Because there is no horizontal band in residuals versus fitted values plot above, we cannot say that our variance is constant. On the contrary, there seems to be an increasing variance problem.

3. **Independence of Errors**

As it can be observed from the above residuals versus observation order plot, residuals bounce randomly around the 0 line which means the independence of errors assumption is valid.

4. **Linearity**

According to the residuals versus fitted values plot above, the residuals do not bounce randomly around 0 line. Thus, we can say that the linearity assumption is not satisfied.

5. **Multicollinearity**

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-25989	10220	-2,54	0,012	
peak-rpm	-0,388	0,652	-0,60	0,552	1,44
horsepower	167,0	12,9	12,92	0,000	3,45
wheel-base	319,6	73,0	4,38	0,000	2,90
height	-3	145	-0,02	0,981	1,86
city-mpg	-38,0	82,9	-0,46	0,647	4,19
FuelType					
gas	-6415	1217	-5,27	0,000	1,97
Aspiration					
turbo	-3951	798	-4,95	0,000	1,39

According to the table above, there does not seem a VIF associated with the parameter estimates bigger than 10 or 5. So we can say that we have no multicollinearity problem in our model.

6. **Outliers and Influential Observations**

- There are three observations that have their standardized residuals' absolute value greater than 3, which are 17th, 73rd and 75th observation. They can be considered as outliers.
- Leverage values are compared with $\frac{2p}{n} = \frac{2.8}{205} = 0.078$ and there are twenty observations that have a value greater than 0.078. They are 7th, 8th, 9th, 48th, 49th, 50th,

67th, 71st, 73rd, 74th, 91st, 105th, 106th, 111st, 115th, 127th, 128, 129th, 159th and 160th observations. These observations seem to be leverage points.

- In our data, because n>3p (205 > 3.8), the interval for the condition that an observation being an influential point is being smaller than $\frac{1-3p}{n}$ or being larger than $\frac{1+3p}{n}$. For our data, this interval is computed as being smaller than 0.112 and being larger than 0.1219. The observations that have this condition are 73rd and 129th observations.
- When DFFITS values are compared with threshold of $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{8}{205}} = 0.3950$, there are thirteen observations that have a larger value than 1.13137. They are 17th, 19th, 49th, 70th, 71st, 72nd , 73rd , 74th, 75th, 127th, 128th, 129th and 205th values.

CONCLUSION

Above residual analysis indicates that residuals do not follow a normal distribution, there seems to be an increasing variance problem and linearity assumption is not satisfied. On the other hand, there are three outliers (17th, 73rd and 75th) other than influential points. There are two observations (73rd and 129th) that are detected to be influential by Cook's Distance and thirteen (17th, 19th, 49th, 70th, 71st, 72nd , 73rd , 74th, 75th, 127th, 128th, 129th and 205th) observations that are detected to be influential by DFFITS.

APPENDIX 1: MINITAB Output

Regression Analysis: price versus horsepower; peak-rpm; wheel-base; height; width; Fuel Type; Aspiration

Method

Categorical predictor coding	(1; 0)
Rows unused	6

Regression Equation

FuelType	Aspiration	
diesel	std	price = -69049 + 153,4 horsepower - 0,330 peak-rpm + 115,0 wheel-base + 87 height + 876 width
diesel	turbo	price = -72933 + 153,4 horsepower - 0,330 peak-rpm + 115,0 wheel-base + 87 height + 876 width
gas	std	price = -74522 + 153,4 horsepower - 0,330 peak-rpm + 115,0 wheel-base + 87 height + 876 width
gas	turbo	price = -78407 + 153,4 horsepower - 0,330 peak-rpm + 115,0 wheel-base + 87 height + 876 width

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-69049	14516	-4,76	0,000	
horsepower	153,4	10,4	14,71	0,000	2,37
peak-rpm	-0,330	0,632	-0,52	0,602	1,43
wheel-base	115,0	93,8	1,23	0,221	5,04
height	87	144	0,60	0,548	1,93
width	876	269	3,25	0,001	4,99
FuelType					
gas	-5474	1124	-4,87	0,000	1,77
Aspiration					
turbo	-3885	772	-5,03	0,000	1,37

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3578,90	80,59%	79,88%	78,13%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	10158213502	1451173357	113,30	0,000
horsepower	1	2770835453	2770835453	216,33	0,000
peak-rpm	1	3487526	3487526	0,27	0,602
wheel-base	1	19275413	19275413	1,50	0,221
height	1	4633710	4633710	0,36	0,548
width	1	135545524	135545524	10,58	0,001
FuelType	1	303992241	303992241	23,73	0,000
Aspiration	1	323942593	323942593	25,29	0,000
Error	191	2446421968	12808492		
Lack-of-Fit	119	2335766196	19628287	12,77	0,000
Pure Error	72	110655773	1536886		
Total	198	12604635471			

Fits and Diagnostics for Unusual Observations

Obs	price	Fit	Resid	Std Resid	
17	41315	27659	13656	3,89	R
50	36000	41742	-5742	-1,74	X
73	35056	24963	10093	3,01	R X
74	40960	33853	7107	2,10	R
75	45400	32979	12421	3,60	R
102	13499	21661	-8162	-2,31	R
103	14399	21748	-7349	-2,09	R
104	13499	21661	-8162	-2,31	R
127	32528	26995	5533	1,65	X
128	34028	26995	7033	2,10	R X
129	37028	26995	10033	2,99	R X
179	16558	24112	-7554	-2,14	R
180	15998	24112	-8114	-2,30	R
205	22625	15020	7605	2,20	R

R Large residual
X Unusual X