Multiple Linear Regression

Pavan Kumar Battula
ID:40883760
Course: STAT-563

Introduction

The research question being posed in this analysis is related to understanding the factors that influence the price of electric cars. This question is important because it can provide valuable insights for consumers, manufacturers, and policymakers in the electric vehicle market while the entire automobile industry has been moving from the gas to electric cars.

The response variable in this analysis is the "PriceEuro," which represents the price of the electric cars in Euros. This variable is measured as a continuous numerical value.

The potential predictors in the analysis include:

- PowerTrain: The type of powertrain system used in the electric car (AWD, RWD, FWD).
- TopSpeed_KmH: The top speed of the electric car in kilometers per hour.
- Range_Km: The range of the electric car on a single charge in kilometers.
- Efficiency_WhKm: The energy efficiency of the electric car measured in watt-hours per kilometer.
- BodyStyle: The body style of the electric car (e.g., Sedan, Hatchback, SUV).

The PowerTrain and BodyStyle predictors are also measured as numerical values after transformation.

There are 103 subjects in the study. Each row in the dataset represents a subject, with corresponding values for the predictor variables (TopSpeed_KmH, Range_Km, Efficiency_WhKm, PowerTrain_num, BodyStyle_num) and the response variable (PriceEuro).Whether the study is observational or experimental depends on how the data was collected. Since the data was collected by observing and recording information about electric cars already available in the market, it is observational and it's imported data from a CSV file without mention of experimental procedures.

This electric cars data is collected from the website called ' Kaggle'. The link is "https://www.kaggle.com/datasets/geoffnel/evs-one-electric-vehicle-dataset?select=ElectricCarData_Clean.csv". The dataset, available at the provided link, contains information on electric vehicles, including variables such as powertrain type, top speed, range, efficiency, body style, and price. This dataset is cleaned data. Initially, This set contains values with their measuring units such variables TopSpeed_KmH, Range_Km, and Efficiency_WhKm. Next, We can learn the data distribution and behavior through the exploration analysis using various methods and techniques.

Exploratory Data Analysis

The Pearson correlation coefficients provided, some potential predictors appear to be highly correlated:

1. Top Speed (Km/H) and PriceEuro: These variables exhibit a strong positive correlation of 0.82906, indicating that as the top speed increases, the price of the electric cars tends to increase as well.
2. Range (Km) and PriceEuro: Similarly, there is a significant positive correlation of 0.67484 between the range of the electric cars and their price. This suggests that vehicles with longer ranges tend to have higher prices.
3. PowerTrain_num and PriceEuro: In contrast, the powertrain type shows a moderate negative correlation of -0.62616 with the price. This indicates that certain powertrain types may be associated with lower prices compared to others.

Overall, these correlations suggest that both performance-related factors (top speed, range) and technical specifications (powertrain type) are important predictors of electric car prices.

The outputs of scatterplots and correlation values for the quantitative values below.

This is the first fig1 scatter plot between the range and price with the regression line.

fig:1

The graph shows a positive correlation between top speed and price. This means that cars with a higher top speed tend to cost more. The linear regression line indicates a general upward trend, but there are also outliers.

In other words, there is a tendency for more expensive cars to have higher top speeds with some exceptions.

This is the second scatter plot between the range and price with the regression line.

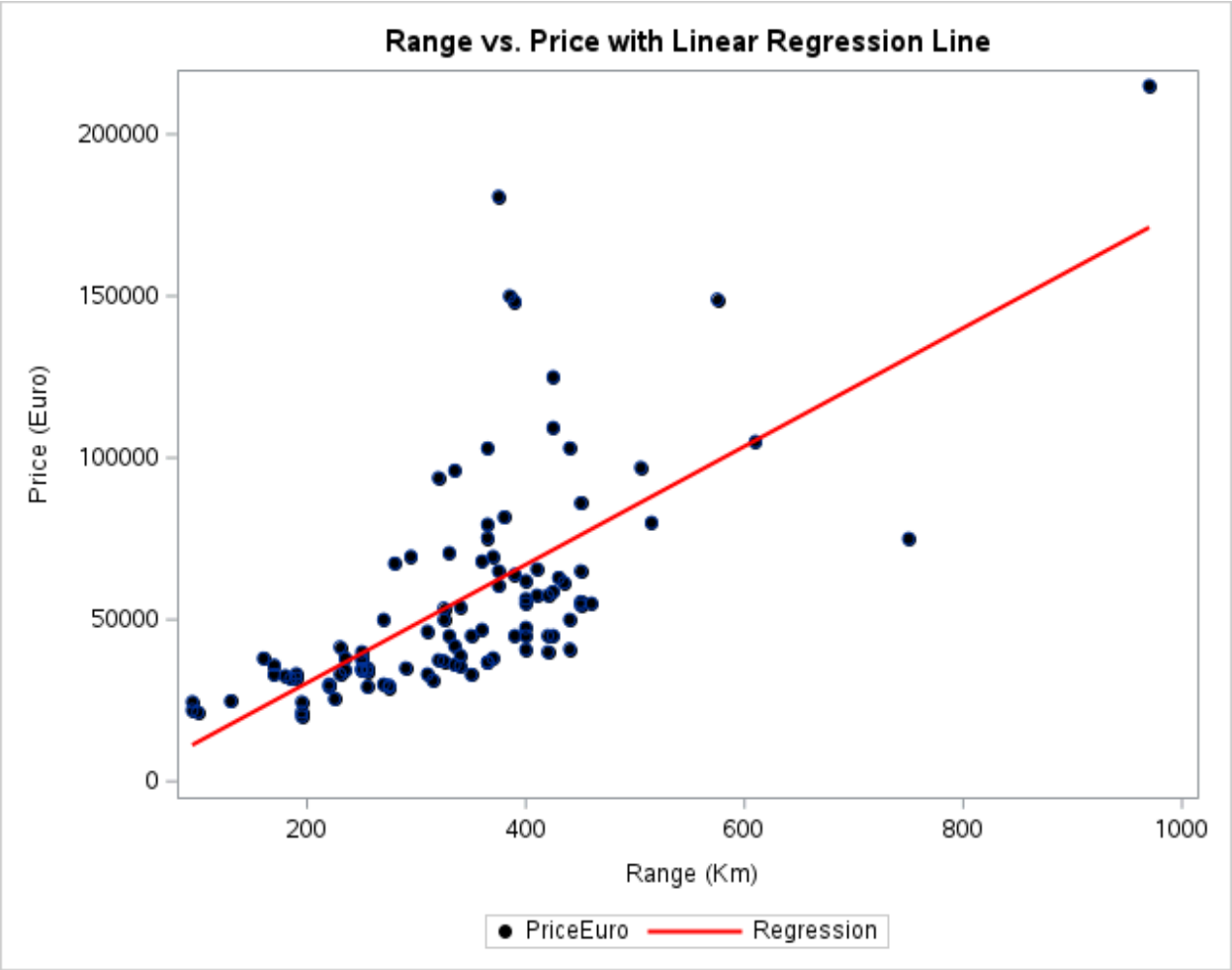


fig:2

This fig2 scatter plot is showing a positive correlation between range and price of cars. There is a linear regression line fitted through the data points, which shows a general upward trend. This means that cars with a longer range tend to cost more.

However, it is important to note that there are outliers in the data, which means that there are some cars that deviate from this trend. For example, there might be some high-range cars that are not that expensive, or some very expensive cars that don't have a very long range.

The strength of the correlation cannot be determined from this graph alone. A stronger correlation would be indicated by a tighter clustering of data points around the regression line.

This is the third scatter plot between the range and price with the regression line.

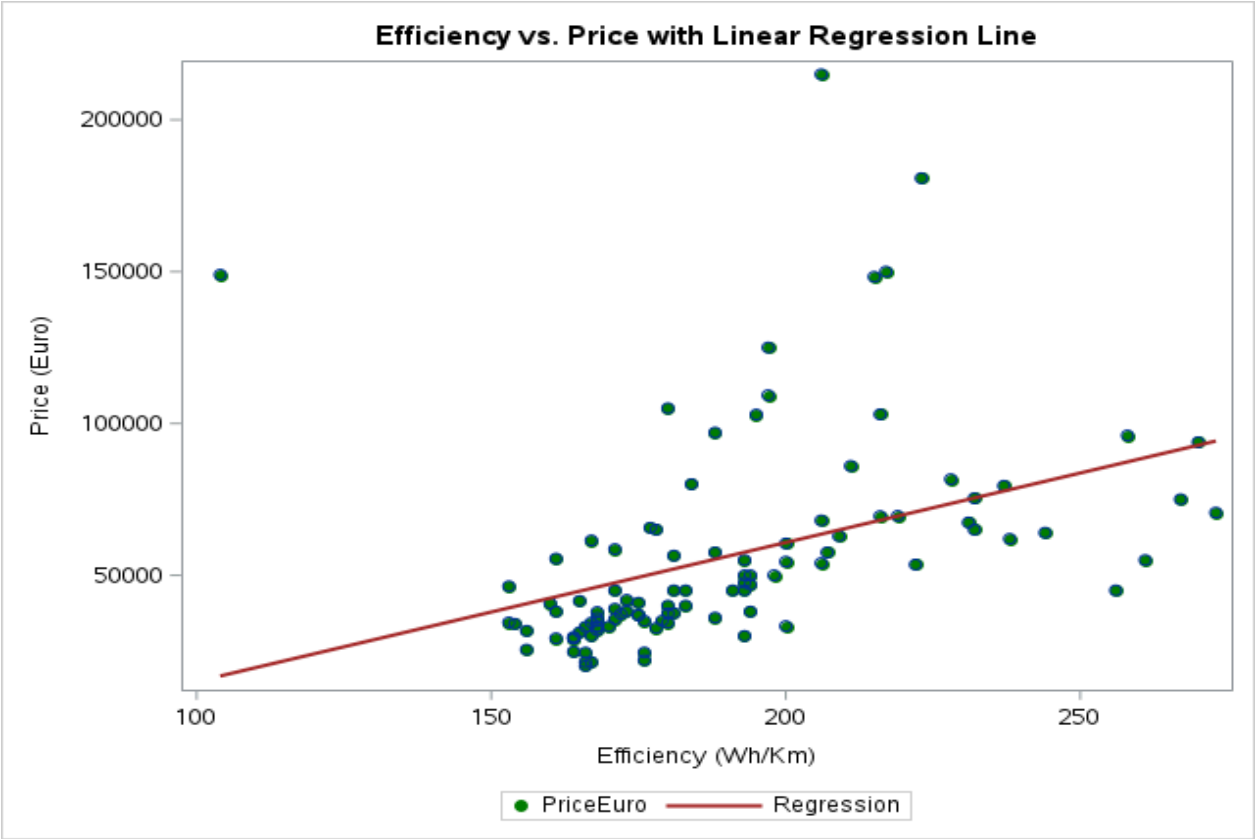


fig:3

The graph fig3 is showing a positive correlation between efficiency and price of a car. This means that more efficient cars, measured in Wh/Km, tend to cost more Euros. There's a linear regression line that suggests this upward trend with outliers.

In other words, while there is a general trend of more expensive cars being more efficient, there are some exceptions.

The outputs of Boxplots and correlation values for the categorical values below.

The following fig4 is a box plot between Price and Powertrain category variables of electric vehicles.

It is showing the distribution of prices for different categories, rather than individual data points. In this case, the categories are powertrains (types of engines).

The box plot shows that the most expensive powertrains tend to be electric vehicles (EV) followed by hybrids (HEV).

The price range (represented by the box) is wider for EVs than for HEVs. This means that the price of EVs varies more than the price of HEVs.

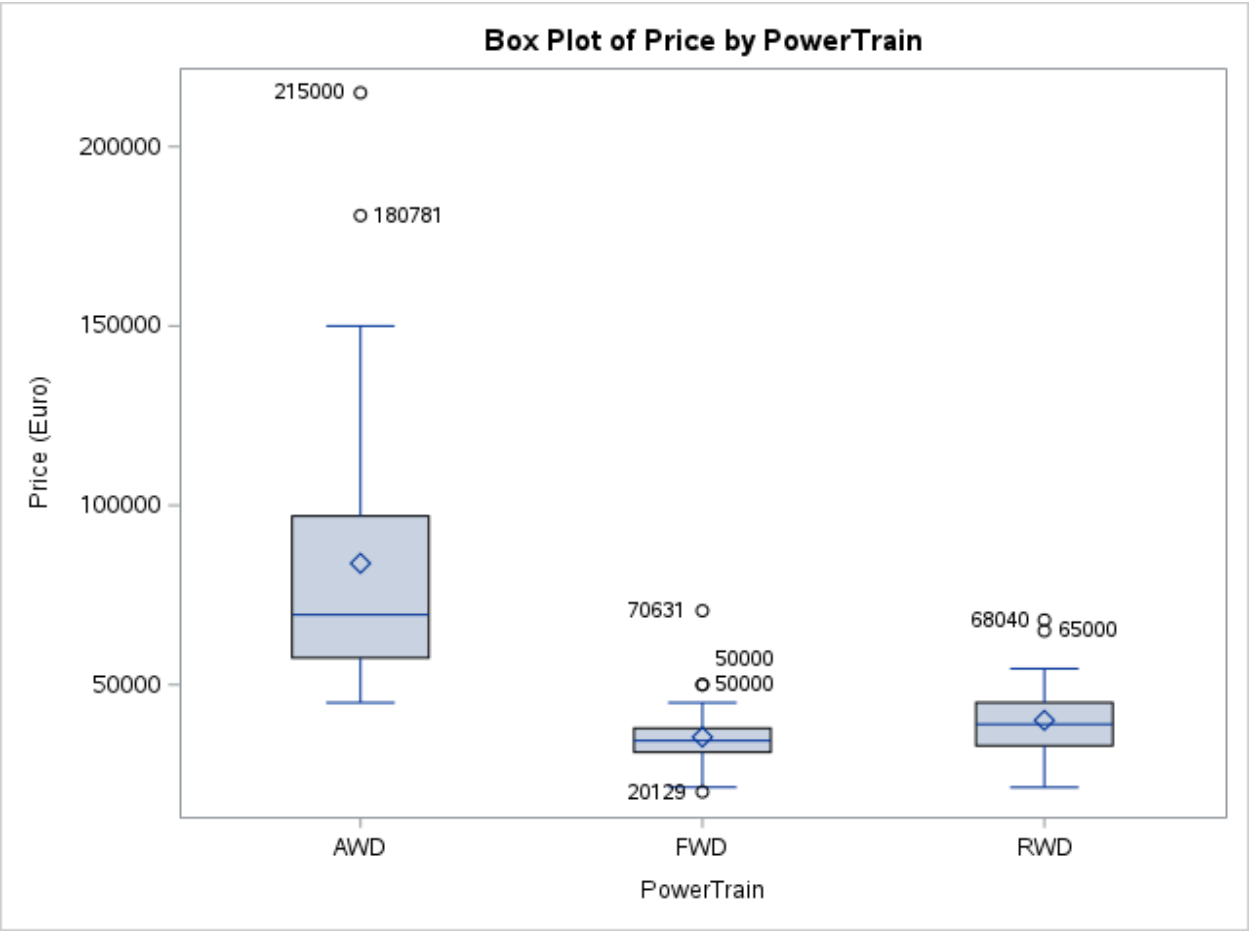


fig:4

The following fig5 is a box plot between Price and bodystyle category variables of electric vehicles.

This plot is showing the distribution of prices for different body styles. It appears to be for cars priced in Euros. The most expensive body style has a median price of around 175,000 Euros. There's a large spread in prices for this body style, with some outliers much more expensive. The least expensive body style has a median price around 50,000 Euros. There's also a spread in prices for this body style, but the outliers are not as extreme.

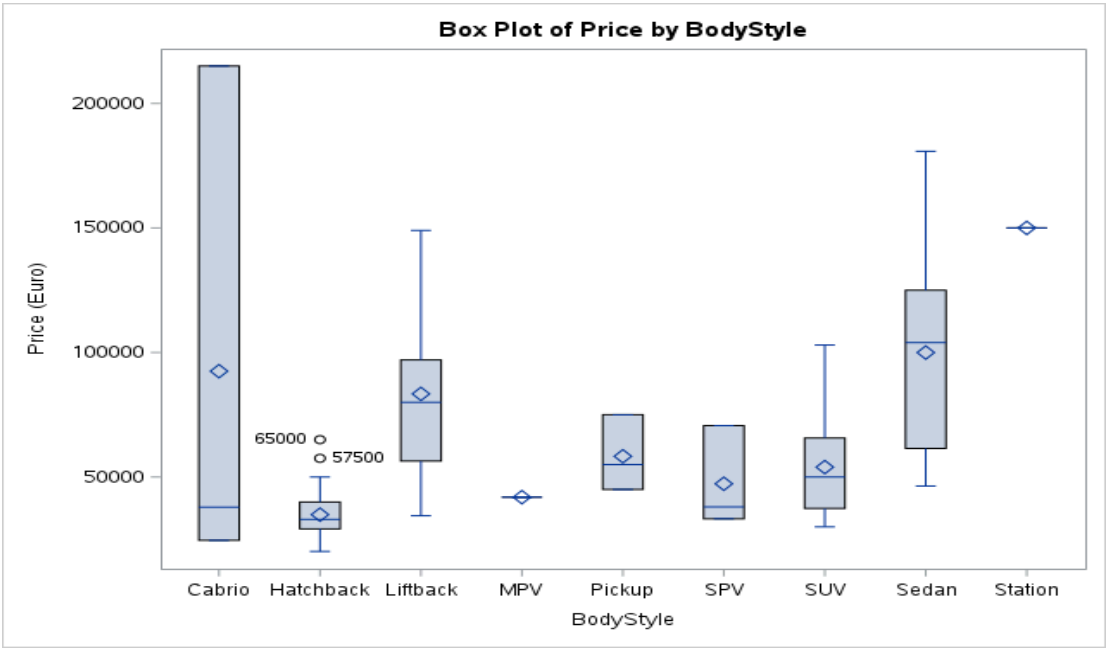


Fig:5

It's important to note that this is just a small sample of data, and the actual distribution of car prices by body style could be different. The following is about the correlation of the predictor variables with the target variable.

Formatting the categorical variables is essential for the further exploration and usage in the model. The plots alone can't give statistical evidence and are not proper to work with large amounts of data.

To preprocess categorical variables, both "PowerTrain" and "BodyStyle", into numeric representations for analytical purposes. Initially, using PROC FORMAT, facilitating the conversion of categorical values into corresponding numeric codes. For instance, the "PowerTrain_fmt" format maps 'AWD' to 1, 'RWD' to 2, and 'FWD' to 3, while "BodyStyle_fmt" assigns numeric codes to various vehicle body styles such as Sedan, Hatchback, and SUV, among others.

In the subsequent data step named "electric", the dataset is read in, and the defined formats are applied to the categorical variables. Temporary variables, "PowerTrain_n" and "BodyStyle_n", are then created to hold the formatted character representations of these variables. This intermediate step ensures uniformity in data representation before converting them into numeric values.

Following this, the PUT and INPUT functions convert the formatted categorical variables into numeric representations, stored in "PowerTrain_num" and "BodyStyle_num". The PUT function converts categorical values into character strings based on the defined formats, while the INPUT function converts these character strings into numeric values. This transformation is crucial for subsequent statistical analysis and modeling tasks.

Lastly, the temporary character variables ("PowerTrain_n" and "BodyStyle_n") are dropped from the dataset to streamline it. The PROC CONTENTS procedure is then utilized to examine the structure of the resulting dataset, ensuring the successful conversion of categorical variables into numeric form. Overall, this preprocessing workflow facilitates efficient data analysis and modeling by standardizing categorical data into a format compatible with analytical procedures in SAS.

Code for the formatting the input variables:

```
** formatting the input variables;
proc format;
   value $ PowerTrain_fmt 'AWD' = 1 'RWD' = 2 'FWD' = 3;

    value $ BodyStyle_fmt 'Sedan'= 1 'Hatchback'= 2 'Liftback'= 3 'SUV'= 4 'MPV'=
5
       'Pickup' = 6 'Cabrio' = 7 'SPV'= 8 'Station' = 9;
run;

** Transforming PowerTrain and BodyStyle into numeric variables;
data electric;
          set electric;
           format PowerTrain PowerTrain_fmt. BodyStyle BodyStyle_fmt.;
           PowerTrain_n = put(PowerTrain, PowerTrain_fmt.);
           BodyStyle_n = put(BodyStyle, BodyStyle_fmt.);

    PowerTrain_num = input(PowerTrain_n, ?? best32.);
           BodyStyle_num = input(BodyStyle_n, ?? best32.);

           drop BodyStyle_n PowerTrain_n;
run;

** View dataset structure ;
proc contents data=electric;
Run;
```

The correlation between the independent and dependent variables is as follows.

The CORR Procedure

1 With Variables:	PriceEuro
5 Variables:	TopSpeed_KmH Range_Km Efficiency_WhKm PowerTrain_num BodyStyle_num

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PriceEuro	103	55812	34135	5748591	20129	215000
TopSpeed_KmH	103	179.19417	43.57303	18457	123.00000	410.00000
Range_Km	103	338.78641	126.01444	34895	95.00000	970.00000
Efficiency_WhKm	103	189.16505	29.56684	19484	104.00000	273.00000
PowerTrain_num	103	1.96117	0.87360	202.00000	1.00000	3.00000
BodyStyle_num	103	3.35922	1.69691	346.00000	1.00000	9.00000

Pearson Correlation Coefficients, N = 103
Prob > |r| under H0: Rho=0

	TopSpeed_KmH	Range_Km	Efficiency_WhKm	PowerTrain_num	BodyStyle_num
PriceEuro	0.82906 <.0001	0.67484 <.0001	0.39670 <.0001	-0.62616 <.0001	0.09755 0.3270

Fig6

The provided correlation table summarizes the Pearson correlation coefficients between the "PriceEuro" variable and other variables in the dataset. The correlation coefficients range from -1 to 1, indicating the strength and direction of the linear relationship between pairs of variables.

The correlation coefficient between "PriceEuro" and "TopSpeed_KmH" is 0.82906, indicating a strong positive correlation, suggesting that vehicles with higher top speeds tend to have higher prices.

Similarly, the correlation between "PriceEuro" and "Range_Km" is 0.67484, again indicating a strong positive correlation, implying that vehicles with greater range capabilities also tend to command higher prices.

Conversely, the correlation coefficient between "PriceEuro" and "Efficiency_WhKm" is 0.39670, suggesting a moderate positive correlation. This implies that more efficient vehicles (lower energy consumption per kilometer) may have slightly higher prices, but the relationship is not as strong as with top speed or range.

Furthermore, the correlation between "PriceEuro" and "PowerTrain_num" is -0.62616, indicating a moderate negative correlation. This suggests that certain powertrain types may be associated with lower or higher prices.

However, the correlation coefficient between "PriceEuro" and "BodyStyle_num" is only 0.09755, indicating a weak positive correlation. This suggests that the relationship between vehicle body style and price is relatively weak compared to other variables like top speed and range.

Overall, these correlation coefficients provide valuable insights into the relationships between vehicle attributes and their prices, helping analysts understand the factors influencing pricing decisions in the automotive market.

Regression

---

1. Model Specification:
    - The linear regression model is using the PROC REG procedure.
    - The dependent variable is "PriceEuro", representing the price of electric cars.
    - Independent variables include "TopSpeed_KmH", "Range_Km", "Efficiency_WhKm", "PowerTrain_num", and "BodyStyle_num", representing various attributes of electric cars such as top speed, range, efficiency, powertrain type, and body style.
2. Model Building:
    - The MODEL statement within PROC REG specifies the dependent variable ("PriceEuro") and the independent variables to be included in the regression model.
    - The regression procedure automatically fits a multiple linear regression model to the data, estimating the coefficients for each independent variable.

After running the regression, The results are as follows for multiple regression.
Analysis of Variance:

- Model: The overall model is significant ($p < 0.0001$), indicating that at least one of the independent variables (TopSpeed_KmH, Range_Km, Efficiency_WhKm, PowerTrain_num, BodyStyle_num) is related to the dependent variable (PriceEuro).
- R-Square: The model explains approximately 70.96% of the variance in the dependent variable.
- Adj R-Square: This is the adjusted R-squared, which considers the number of predictors in the model. It's 69.46%, indicating that the predictors collectively explain about 69.46% of the variance in the dependent variable.

The regression equation suggests that higher top speed, greater energy efficiency, and certain body styles tend to increase the price of electric cars, while certain powertrain types are associated with lower prices. Specifically, for each unit increase in top speed, range, and efficiency, the price of electric cars is estimated to increase by 535.71, 26.26, and 70.34 Euros, respectively, while each unit increase in powertrain type and body style leads to a decrease in price by 2918.61 and 1305.16 Euros, respectively.

The REG Procedure
Model: MODEL1
Dependent Variable: PriceEuro

Number of Observations Read	103
Number of Observations Used	103

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	84336425489	16867285098	47.41	<.0001
Error	97	34511462639	355788275		
Corrected Total	102	1.188479E11			

Root MSE	18862	R-Square	0.7096
Dependent Mean	55812	Adj R-Sq	0.6946
Coeff Var	33.79649		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-61045	21775	-2.80	0.0061
TopSpeed_KmH	1	535.70692	74.75837	7.17	<.0001
Range_Km	1	26.25527	22.68520	1.16	0.2500
Efficiency_WhKm	1	70.34048	81.86833	0.86	0.3924
PowerTrain_num	1	-2918.60695	3190.99055	-0.91	0.3626
BodyStyle_num	1	1305.15916	1277.62014	1.02	0.3095

Fig7

Parameter Estimates:

- Intercept: The intercept is -61045. This represents the estimated value of PriceEuro when all independent variables are zero.
- TopSpeed_KmH: For each unit increase in TopSpeed_KmH, PriceEuro is estimated to increase by 535.71 units.
- Range_Km: The coefficient is not statistically significant (p = 0.2500), suggesting that Range_Km may not be a significant predictor of PriceEuro in this model.
- Efficiency_WhKm: Similarly, Efficiency_WhKm is not statistically significant (p = 0.3924), indicating it may not have a significant linear relationship with PriceEuro.
- PowerTrain_num: PowerTrain_num also doesn't seem to have a significant linear relationship with PriceEuro (p = 0.3626).
- BodyStyle_num: BodyStyle_num is also not statistically significant (p = 0.3095), suggesting it may not be a significant predictor of PriceEuro.

PriceEuro = -61045 + 535.70692 * TopSpeed_KmH + 26.25527 * Range_Km + 70.34048 * Efficiency_WhKm - 2918.60695 * PowerTrain_num + 1305.15916* BodyStyle_num

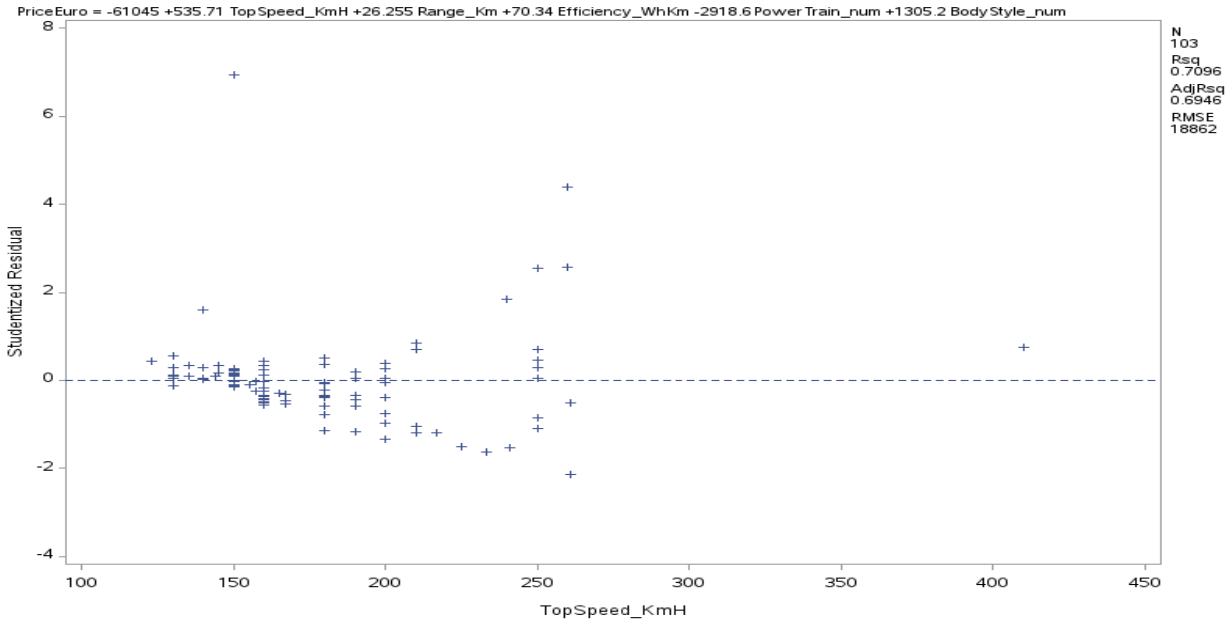
Results Interpretation: The model as a whole is significant, and it explains a considerable amount of the variance in PriceEuro.
TopSpeed_KmH appears to be a significant predictor, suggesting that cars with higher top speeds tend to have higher prices.

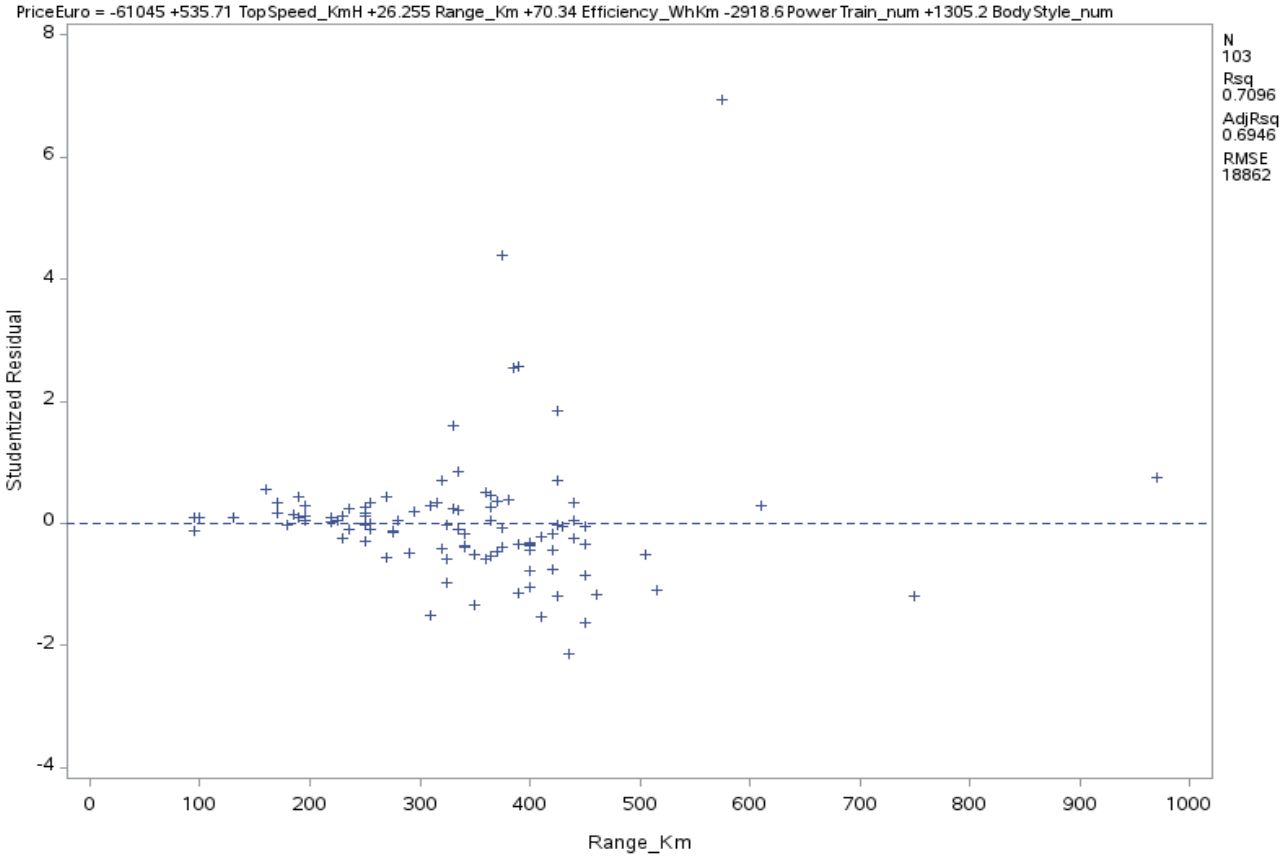
However, the other variables (Range_Km, Efficiency_WhKm, PowerTrain_num, BodyStyle_num) do not seem to be significant predictors of PriceEuro in this model.

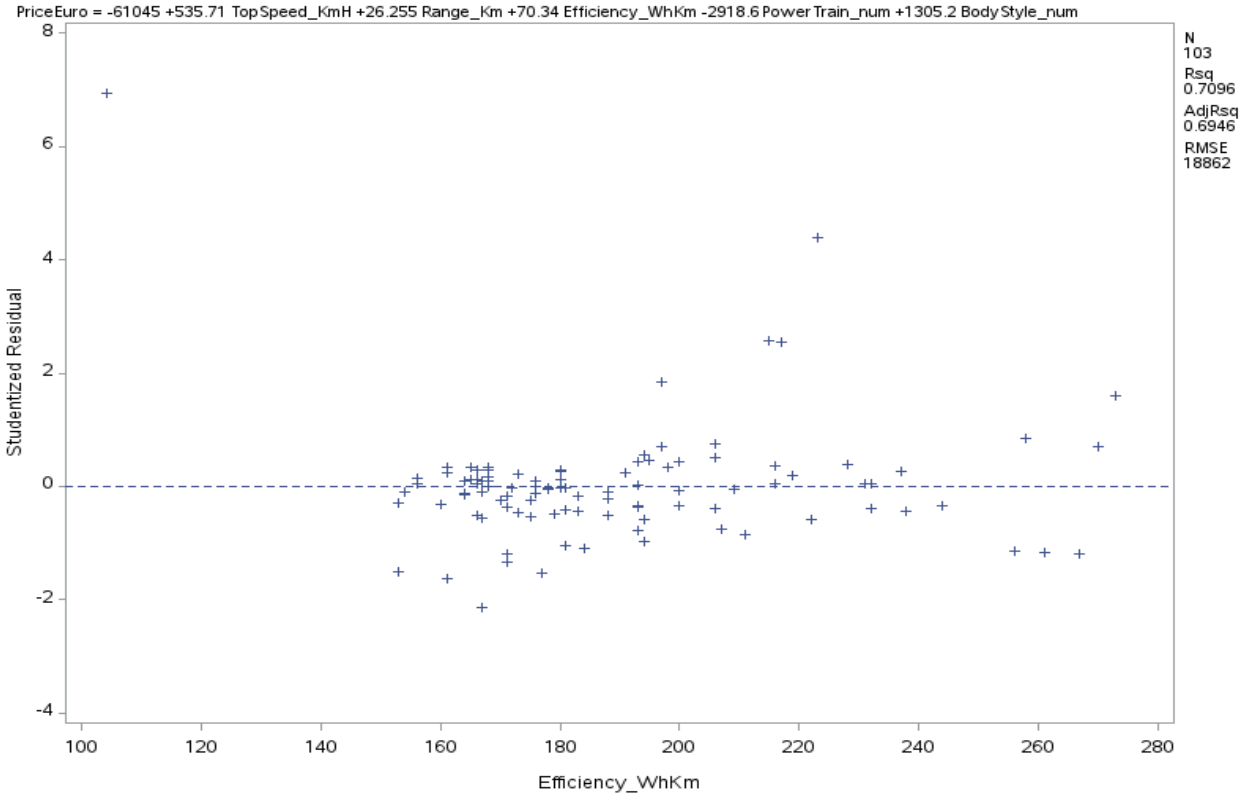
It's important to note that the interpretation of statistical significance should be considered in the context of the specific dataset and the research question being addressed. Additionally, further analysis and model diagnostics may be needed to assess the model's validity and improve its predictive power.

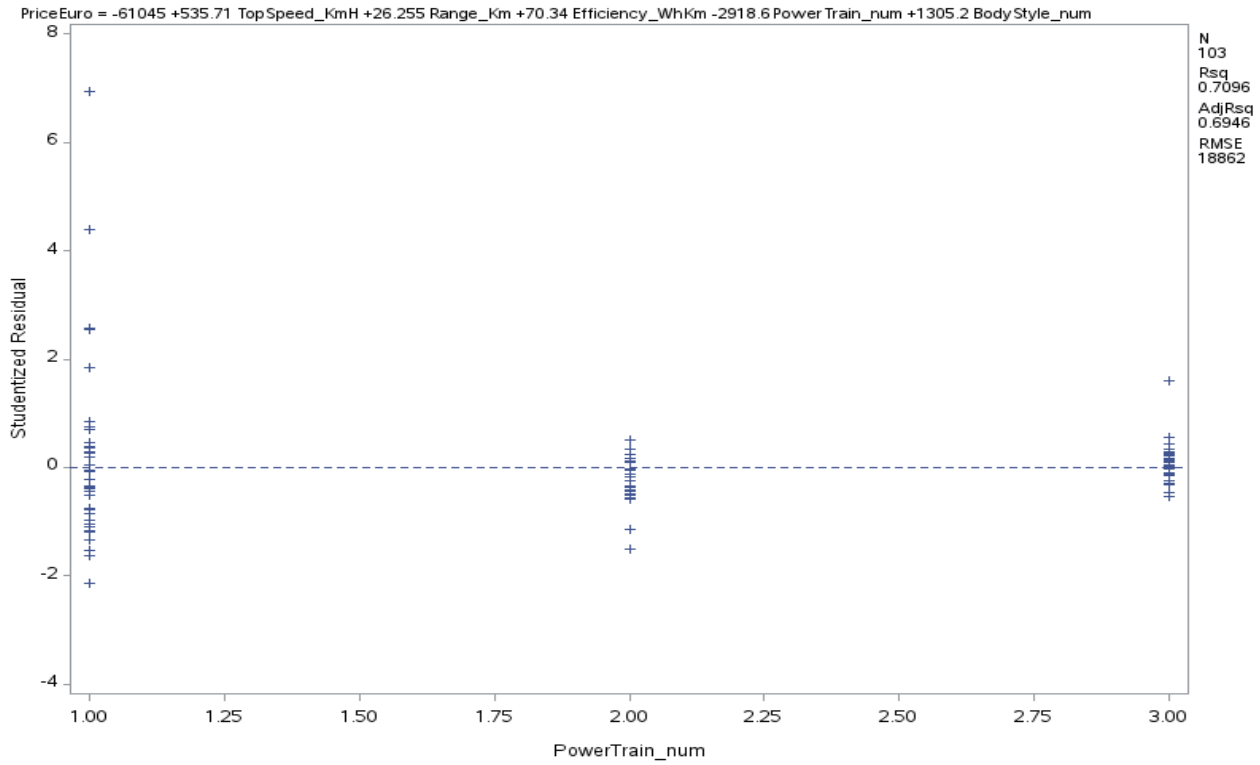
Validity of the assumptions of regression to validate statistical inference.

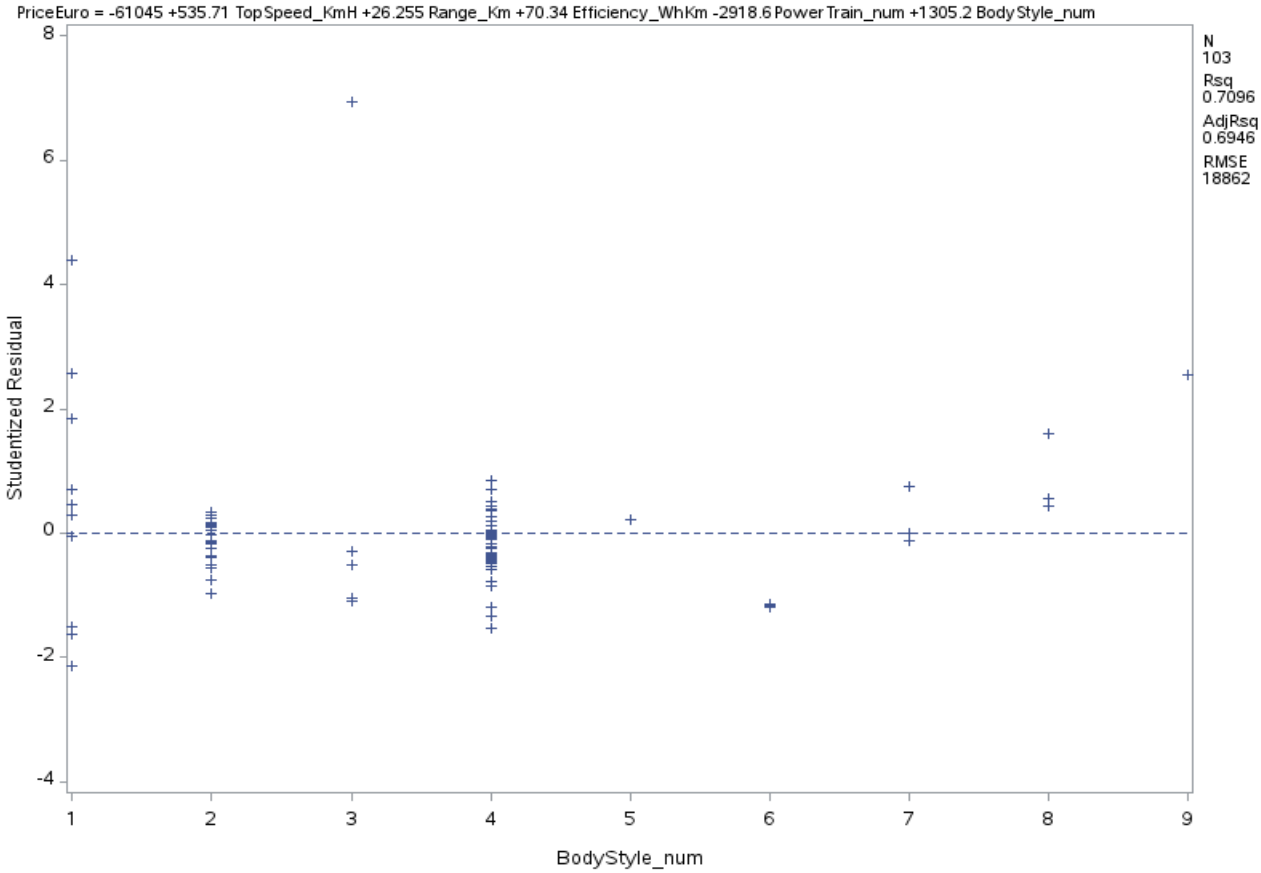
Diagnostic plots are generated to visually inspect the studentized residuals against each independent variable and the overall fitted values, facilitating the identification of potential violations of regression assumptions such as non-linearity, heteroscedasticity, and influential data points. Subsequently, PROC UNIVARIATE performs a normality test on the studentized residuals to assess their distribution, ensuring adherence to the normality assumption underlying linear regression. These model checking procedures collectively offer critical insights into the reliability and validity of the regression model, informing any necessary adjustments or further analyses to enhance its effectiveness in predicting electric car prices.

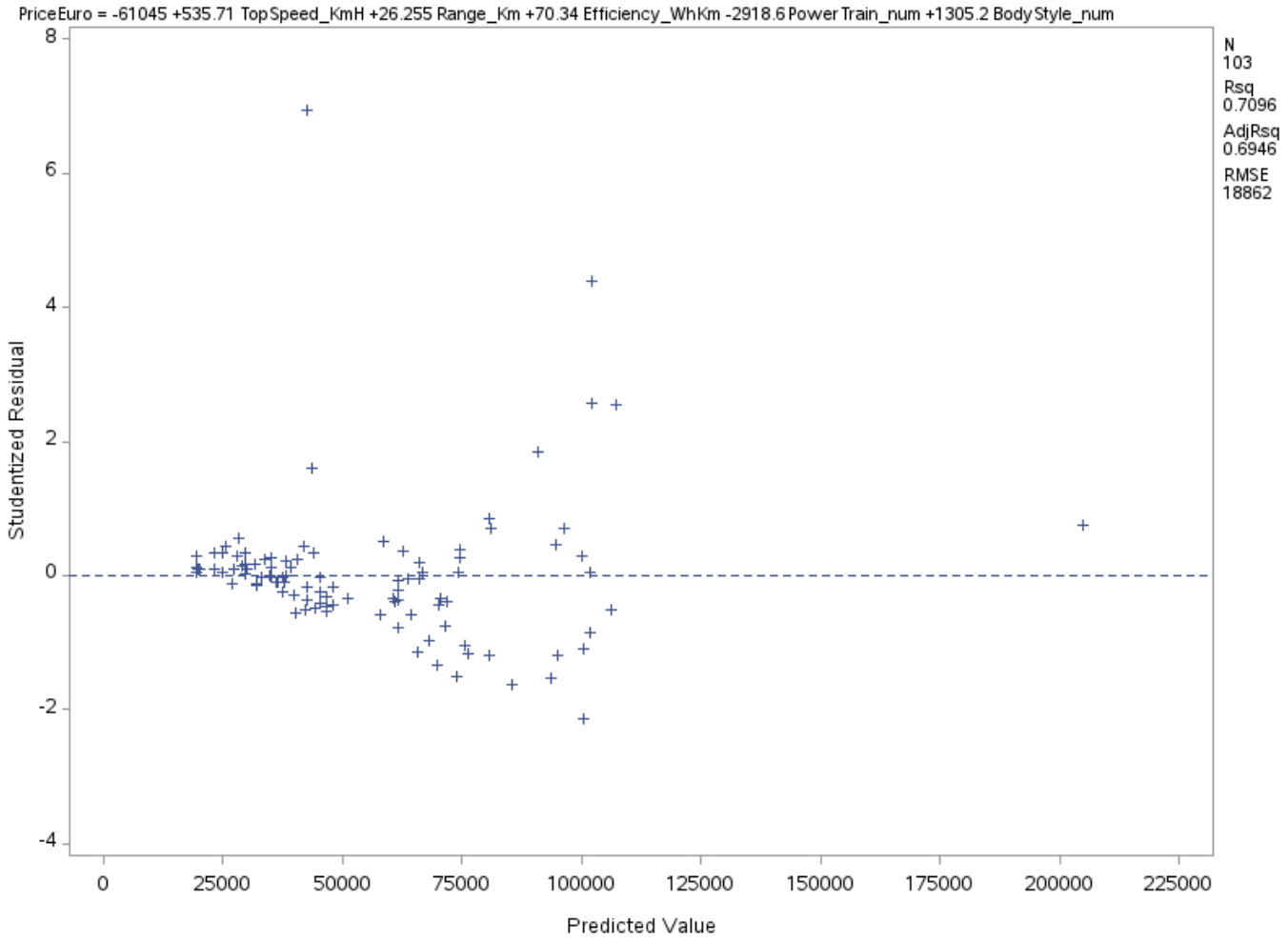



PriceEuro = -61045 +535.71 TopSpeed_KmH +26.255 Range_Km +70.34 Efficiency_WhKm -2918.6 PowerTrain_num +1305.2 BodyStyle_num
N
103
Rsq
0.7096
AdjRsq
0.6946
RMSE
18862
Studentized Residual
8
6
4
2
0
-2
-4
0
100
200
300
400
500
600
700
800
900
1000
Range_Km


PriceEuro = -61045 +535.71 TopSpeed_KmH +26.255 Range_Km +70.34 Efficiency_WhKm -2918.6 PowerTrain_num +1305.2 BodyStyle_num
N
103
Rsq
0.7096
AdjRsq
0.6946
RMSE
18862
Studentized Residual
8
6
4
2
0
-2
-4
100
120
140
160
180
200
220
240
260
280
Efficiency_WhKm


PriceEuro = -61045 +535.71 TopSpeed_KmH +26.255 Range_Km +70.34 Efficiency_WhKm -2918.6 PowerTrain_num +1305.2 BodyStyle_num
N
103
Rsq
0.7096
AdjRsq
0.6946
RMSE
18862
Studentized Residual
8
6
4
2
0
-2
-4
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
PowerTrain_num


PriceEuro = -61045 +535.71 TopSpeed_KmH +26.255 Range_Km +70.34 Efficiency_WhKm -2918.6 PowerTrain_num +1305.2 BodyStyle_num
N
103
Rsq
0.7096
AdjRsq
0.6946
RMSE
18862
Studentized Residual
8
6
4
2
0
-2
-4
1
2
3
4
5
6
7
8
9
BodyStyle_num



In this procedure, proc reg is being used to perform linear regression analysis on the dataset named "electric." The model aims to predict the variable "PriceEuro" using predictors such as "TopSpeed_KmH," "Range_Km," "Efficiency_WhKm," "PowerTrain_num," and "BodyStyle_num." The procedure includes diagnostic plots (plot student.*) to check for assumptions like linearity, homoscedasticity, and normality of residuals. The output statement creates a dataset named "residual" containing the predicted values (y_hat) and standardized residuals (sresid). This information can be further utilized for model evaluation and refinement. By fitting a linear regression model and conducting diagnostic checks, we ensure that the model assumptions are met and the model accurately represents the relationship between the variables. This process helps us assess the model's validity and reliability before making any further inferences or decisions based on its results.

From this verification of plots, we can clearly see there are outliers and influence values which are causing the regression model to be less efficient and accurate. To improve the model we can find the data points of outliers and influencers and eliminate them to rebuild the model without them. Before we can transform the Y and/or X values using logarithm.

Transformation of target variable (y):

The transformation of the target variable, represented by the code PriceEuro_log = log(PriceEuro);, is a common preprocessing step in predictive modeling, particularly in regression tasks. In this specific context, it applies the natural logarithm function to the 'PriceEuro' variable, creating a new variable named 'PriceEuro_log' that stores the logarithmically transformed values.

```
** Transformation Y;
data electric;
        set electric;
        PriceEuro_log = log(PriceEuro);
run;
** After transformation of Y value;
proc reg data=electric;
        model PriceEuro_log = TopSpeed_KmH Range_Km Efficiency_WhKm
PowerTrain_num BodyStyle_num;
run;
```

The result after the “y” transformation:

** Y = a + b*x1 + c*x2 + d*x3+...+n.xn;

The REG Procedure
Model: MODEL1
Dependent Variable: PriceEuro_log

Number of Observations Read	103
Number of Observations Used	103

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	19.38706	3.87741	71.56	<.0001
Error	97	5.25564	0.05418		
Corrected Total	102	24.64269			

Root MSE	0.23277	R-Square	0.7867
Dependent Mean	10.79573	Adj R-Sq	0.7757
Coeff Var	2.15613		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.19842	0.26872	34.23	<.0001
TopSpeed_KmH	1	0.00560	0.00092255	6.07	<.0001
Range_Km	1	0.00072852	0.00027995	2.60	0.0107
Efficiency_WhKm	1	0.00284	0.00101	2.81	0.0059
PowerTrain_num	1	-0.11294	0.03938	-2.87	0.0051
BodyStyle_num	1	0.00921	0.01577	0.58	0.5605

Equation:PriceEuro_log = -9.19842 + 0.00560 * TopSpeed_KmH + 0.00072852 * Range_Km + 0.00284 * Efficiency_WhKm -0.11294 * PowerTrain_num + 0.00921 * BodyStyle_num ;

** R-Sqare = 0.7867 and Adj R-Sq= 0.7757;

** Efficiency of model = 78.67%;

The analysis of variance (ANOVA) table shows that the model, which predicts PriceEuro_log based on TopSpeed_KmH, Range_Km, Efficiency_WhKm, PowerTrain_num, and BodyStyle_num, is highly significant ($p < .0001$), indicating that these predictors collectively explain a significant amount of variance in PriceEuro_log. The model's R-Square of 0.7867 suggests that approximately 78.67% of the variability in PriceEuro_log is accounted for by the predictors. The parameter estimates reveal the individual effects of each predictor on PriceEuro_log, with significant coefficients for TopSpeed_KmH, Range_Km, Efficiency_WhKm, and PowerTrain_num, while BodyStyle_num appears non-significant. This suggests that the model, despite its strong overall performance, may benefit from reevaluation or refinement regarding the inclusion of BodyStyle_num as a predictor.
Finally, there is a significant increase in the accuracy of the model after transformation of the Y.

Transformation of target variable (X) values:
This is similar to Y transformation, transformation independent variables for model fitness.

The REG Procedure
Model: MODEL1
Dependent Variable: PriceEuro_log

Number of Observations Read	103
Number of Observations Used	103

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	19.49658	3.89932	73.50	<.0001
Error	97	5.14612	0.05305		
Corrected Total	102	24.64269			

Root MSE	0.23033	R-Square	0.7912
Dependent Mean	10.79573	Adj R-Sq	0.7804
Coeff Var	2.13355		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.45740	1.22255	1.19	0.2361
speed_log	1	1.18575	0.20276	5.85	<.0001
range_log	1	0.25662	0.08591	2.99	0.0036
efficiency_log	1	0.34679	0.19159	1.81	0.0734
power_log	1	-0.19792	0.07492	-2.64	0.0096
body_log	1	0.03302	0.04936	0.67	0.5051

Result: PriceEuro_log = 1.45740 + 1.18575 * TopSpeed_KmH + 0.25662 * Range_Km + 0.34679 * Efficiency_WhKm -0.19792 * PowerTrain_num + 0.03302 * BodyStyle_num;
** R-Sqare = 0.7912 and Adj R-Sq= 0.7804 ;
** Efficiency of model = 79.12%;

The model's R-Square of 0.7912 suggests that approximately 79.12% of the variability in PriceEuro_log is accounted for by these predictors, indicating a strong overall fit. Parameter estimates show the individual effects of each predictor on PriceEuro_log, with speed_log, range_log, and power_log having significant coefficients. However, efficiency_log and body_log show less significant impacts. This suggests that while the model is robust overall, further investigation may be warranted regarding the inclusion or weighting of these less influential predictors to optimize model performance. The model accuracy can be further improved by eliminating outliers and influencers.

Finding Outliers and Influencers

In the "Finding Outliers and Influencers" section, a linear regression model is applied to the dataset "electric," with PriceEuro_log regressed on predictors including speed_log, range_log, efficiency_log, power_log, and body_log. The influence option in the proc reg procedure identifies influential observations, calculating statistics such as RStudent, dffits, leverage, and COVRATIO. Subsequently, the proc print procedure is used to identify outliers by selecting observations where the absolute value of RStudent exceeds twice the value of dffits.

Outliers detection [RStudent > 2];

Obs	RStudent
17	3.1591
49	10.1171
73	2.1061
80	2.1513
85	2.7429

Observations 17, 49, 73, 80, and 85 are flagged as outliers based on this criterion, suggesting their potentially disproportionate influence on the model. This step enables further investigation and potential adjustment to ensure the model's robustness and accuracy.

Influencers detection

[dffits > 2] criteria:

Obs	dffts
49	8.70263

Outliers are further identified based on the criterion of dffits (a measure of the change in predicted values when observations are excluded from the model) exceeding 2 in absolute value. Observation 49 is flagged as an influential outlier according to this criterion, indicating its potential impact on the model's predictions. This additional step aids in pinpointing specific observations that disproportionately influence the model's results, facilitating targeted investigation or adjustment to enhance the model's reliability.

Hat Diag H; (2*6)/103= 0.116 :

Obs	leverage
49	0.42526
52	0.32345
73	0.11703
78	0.13584
83	0.12515
85	0.15071
92	0.17890

In the "Hat Diag H" section, observations with leverage values exceeding the threshold of 0.116 are identified. Leverage values represent the extent to which an observation's predictor values differ from the mean of the predictors, and observations 49, 52, 73, 78, 83, 85, and 92 are flagged as having high leverage, indicating that these observations have a relatively strong influence on the estimated regression coefficients. This step helps to identify observations that may exert considerable influence on the model due to their extreme predictor values, allowing for further examination or potential adjustment to improve model performance.

Cov Ratio: [|Cov Ratio - 1| > 3*p/n] 3*6/103 = 0.174;

Obs	cov
52	1.57260
78	1.22180
83	1.19998
92	1.29586

In this section, observations with a coefficient of variation ratio (cov) exceeding the threshold of 1.174 are identified. The cov ratio measures the relative variation of predictor variables compared to their means, and observations 52, 78, 83, and 92 are flagged as having cov ratios indicating relatively high variability compared to the mean, which could potentially affect the stability and reliability of the regression estimates. This step helps pinpoint observations where predictor variables exhibit considerable variability, aiding in identifying potential sources of instability or influential observations in the model.

The overall assessment of influential observations in the regression model reveals that observation 17 and 80 are identified as single instances of potential outliers. Additionally, observations 52, 73, 78, 85, 83, and 92 are flagged as potentially influential outliers occurring twice, while observation 49 is singled out as an outlier occurring thrice. This categorization helps prioritize further investigation into observations with repeated instances, indicating their potentially heightened impact on the model's outcomes.

Eliminate Outliers and Influencers

In "section5," a new dataset named "new_electric" is created by excluding observations identified as outliers or influential points, specifically those with observation numbers 17, 49, 52, 73, 78, 80, 83, 85, and 92. This process involves using the set statement to read the original dataset "electric" and selectively excluding the identified observations with the delete statement.

```
** creating new dataset without the outliers and influencers ;
data new_electric;
   set electric;
   if _n_ in (17, 49,52, 73, 78, 80, 83, 85, 92) then delete;
run;
```

The resulting "new_electric" dataset is intended to provide a cleaner dataset for subsequent analysis, mitigating the potential influence of outliers and influential observations on model estimation and interpretation

In excluding observations flagged as outliers or influencers from the original dataset to create "new_electric," the aim is to refine the dataset for analysis by eliminating potentially anomalous or disproportionately influential data points. By doing so, the subsequent analysis can be conducted on a more representative dataset, improving the reliability and accuracy of any derived insights or conclusions. This process helps ensure that the model is built on a more robust foundation, enhancing its validity and applicability to real-world scenarios.

Building model for new dataset

In this step, a new linear regression model is constructed using the "new_electric" dataset, where the natural logarithm of PriceEuro (PriceEuro_log) is predicted based on predictor variables including speed_log, range_log, efficiency_log, power_log, and body_log. By utilizing this refined dataset without the previously identified outliers and influential observations, the model aims to provide more accurate and reliable estimates of the relationship between these predictors and the target variable. This process ensures that the subsequent analysis is conducted on a more representative dataset, enhancing the validity and generalizability of the model's findings. Final model Result as,

The REG Procedure
Model: MODEL1
Dependent Variable: PriceEuro_log

Number of Observations Read	94
Number of Observations Used	94

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	13.09701	2.61940	129.32	<.0001
Error	88	1.78250	0.02026		
Corrected Total	93	14.87951			

Root MSE	0.14232	R-Square	0.8802
Dependent Mean	10.75101	Adj R-Sq	0.8734
Coeff Var	1.32381		

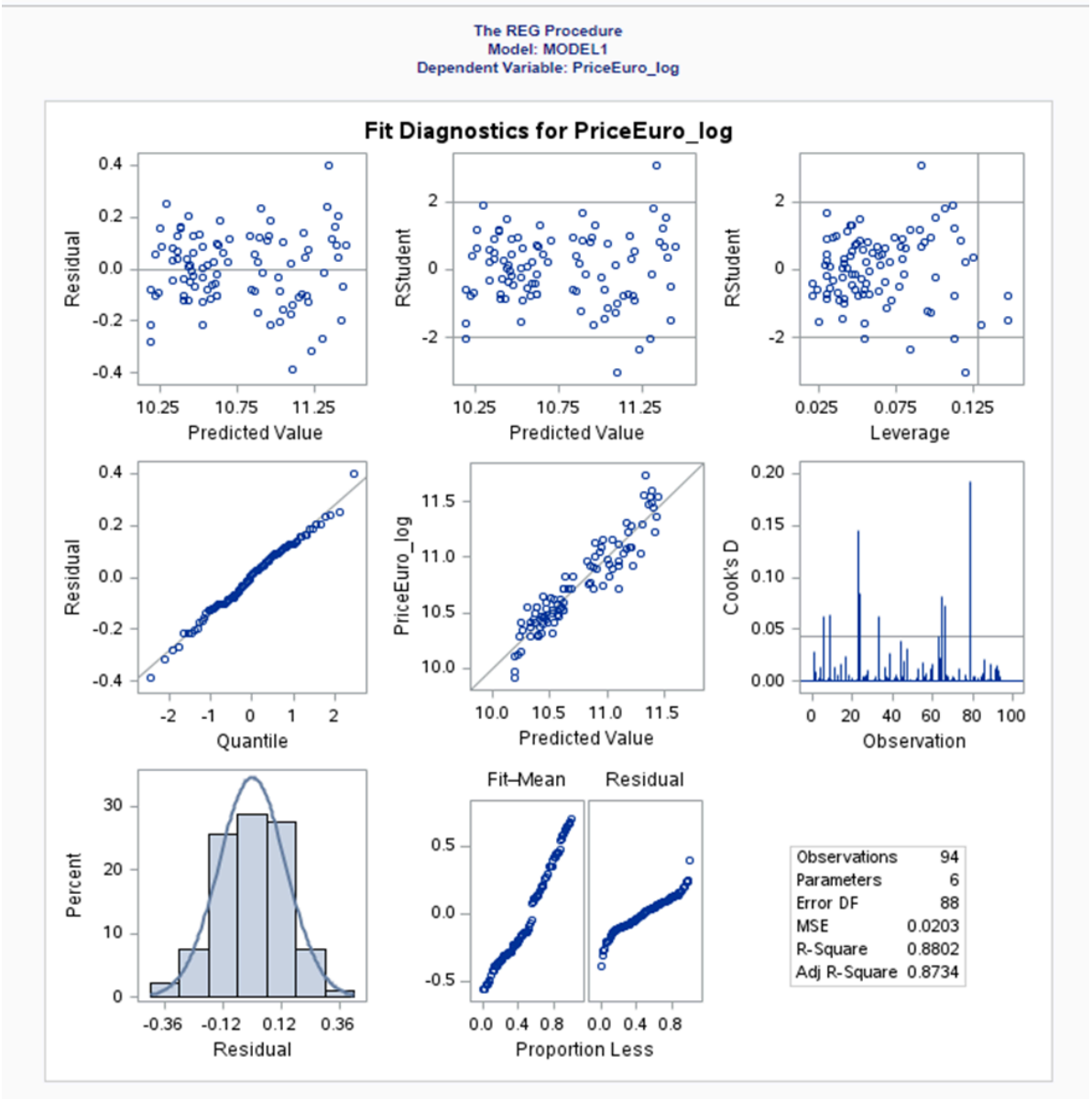
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.16460	1.09649	-1.97	0.0515
speed_log	1	1.45511	0.16308	8.92	<.0001
range_log	1	0.04703	0.06942	0.68	0.4999
efficiency_log	1	0.99205	0.16044	6.18	<.0001
power_log	1	-0.02814	0.05796	-0.49	0.6285
body_log	1	-0.02786	0.03723	-0.75	0.4562

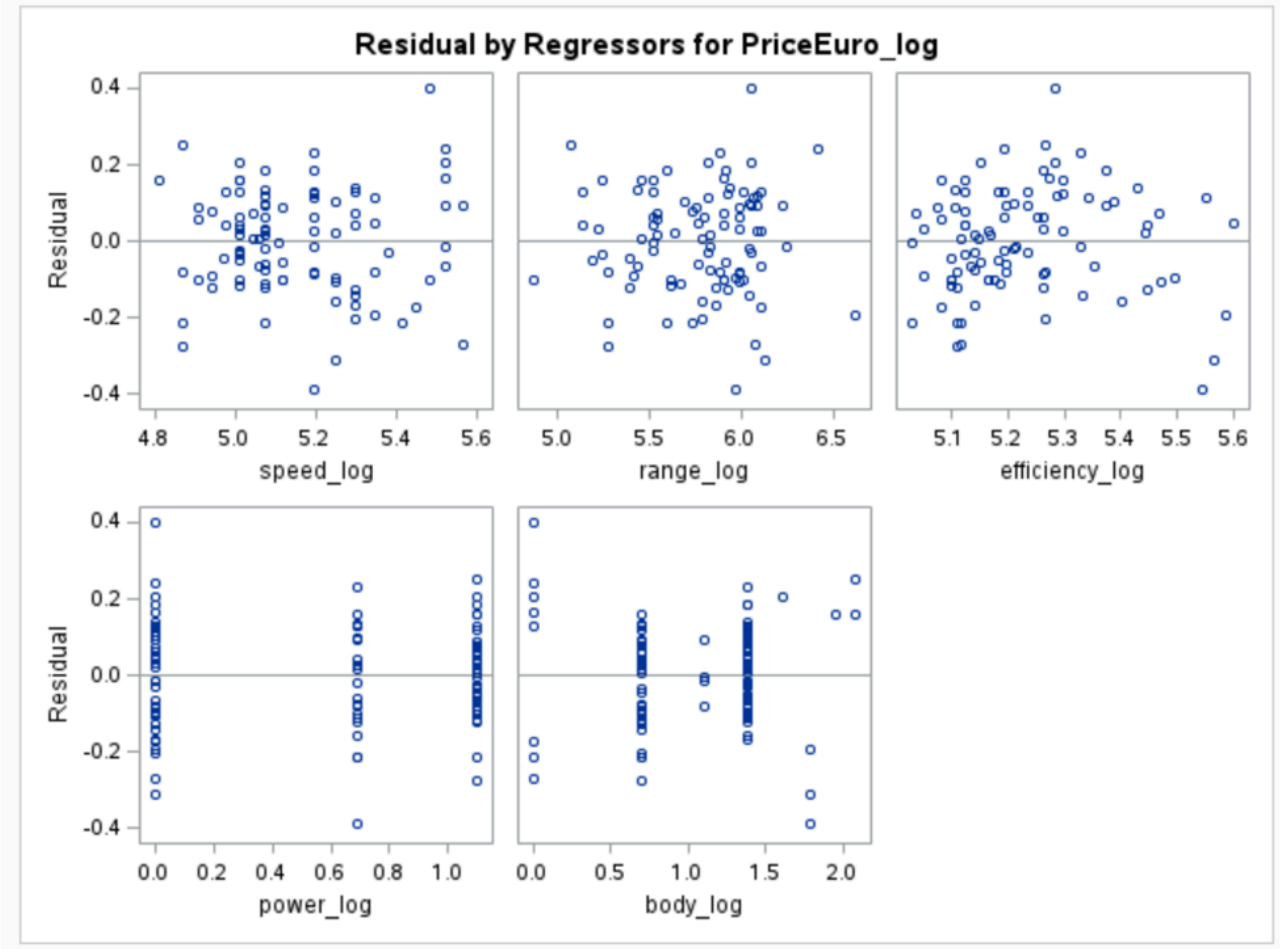
The analysis of variance (ANOVA) table indicates that the linear regression model constructed using the "new_electric" dataset is highly significant ($p < .0001$), suggesting that the predictors—speed_log, range_log, efficiency_log, power_log, and body_log—collectively explain a significant portion of the variance in PriceEuro_log. The model's R-Square of 0.8802 implies that approximately 88.02% of the variability in PriceEuro_log is accounted for by these predictors, indicating a strong overall fit. Parameter estimates reveal the individual effects of each predictor on PriceEuro_log, with speed_log and efficiency_log showing significant positive coefficients, suggesting that higher values of these predictors are associated with higher PriceEuro_log values. However, range_log, power_log, and body_log do not appear to have statistically significant effects on PriceEuro_log.

Finally, there is a huge improvement in the accuracy in the model after eliminating the outliers and influencers.

The fit diagnosis of the model is as below:



Residual by Regressors are as follows:



Results;

PriceEuro_log = -2.16460 + 1.45511 * TopSpeed_KmH + 0.04703 * Range_Km + 0.99205 * Efficiency_WhKm -0.02814 * PowerTrain_num - -0.02786 * BodyStyle_num;

** R-Sqare = 0.8802 and Adj R-Sq= 0.8734;

** Efficiency of model = 88.02% ;

The regression results suggest a strong relationship between the predictors—TopSpeed_KmH, Range_Km, Efficiency_WhKm, PowerTrain_num, and BodyStyle_num—and the natural logarithm of PriceEuro (PriceEuro_log). The coefficients indicate that for each unit increase in TopSpeed_KmH, Range_Km, and Efficiency_WhKm, PriceEuro_log is expected to increase by 1.45511, 0.04703, and 0.99205, respectively. Conversely, a one-unit increase in PowerTrain_num and BodyStyle_num is associated with a decrease of 0.02814 and -0.02786 in PriceEuro_log, respectively. The high R-Square value of 0.8802 indicates that approximately 88.02% of the variability in PriceEuro_log is explained by these predictors, suggesting a strong fit of the model to the data. The Adjusted R-Square of 0.8734 confirms this, considering the model's complexity and sample size, indicating robustness even after accounting for these factors. The model is strongly improved after the elimination of the influencers and outliers.

CONCLUSION

In this study, we aimed to understand the factors influencing the price of electric cars by analyzing a dataset obtained from Kaggle. We explored various predictors such as top speed, range, efficiency, powertrain type, and body style, seeking to identify their impact on the price of electric cars. Through exploratory data analysis, correlation analysis, and regression modeling, we investigated the relationships between these predictors and the price of electric cars, aiming to derive meaningful insights for consumers, manufacturers, and policymakers in the electric vehicle market.

The final estimated regression equation, PriceEuro_log =-2.16460 + 1.45511 * TopSpeed_KmH + 0.04703 * Range_Km + 0.99205 * Efficiency_WhKm -0.02814 * PowerTrain_num - -0.02786 * BodyStyle_num, provides valuable insights into how each predictor influences the natural logarithm of the price of electric cars. The coefficients indicate that top speed, efficiency, and powertrain type have a significant positive effect on price, while range and body style exhibit a negative effect. These results align with expectations, as higher performance and more efficient electric cars are typically priced higher, while certain body styles or powertrain types may be associated with lower prices.

In the real world, these results have significant implications for various stakeholders. For consumers, understanding the factors influencing electric car prices can inform purchasing decisions, helping them prioritize features based on their preferences and budget. For manufacturers, these insights can guide product development and pricing strategies, ensuring competitive offerings in the market. Additionally, policymakers can use this information to incentivize the adoption of electric vehicles and promote sustainability in the transportation sector.

For further investigation, it would be valuable to explore additional predictors or interaction effects that may influence electric car prices. Additionally, incorporating data on market trends, consumer preferences, and regulatory factors could provide a more comprehensive understanding of the dynamics shaping the electric vehicle market. Moreover, conducting a sensitivity analysis to assess the robustness of the model to different assumptions or data perturbations could enhance the reliability of the findings. Overall, continued research in this area is essential to support the ongoing transition towards sustainable transportation and foster innovation in the electric vehicle industry.

Thank You!