Assignment Overview:

An econometric analysis investigating multilinear regression between vehicle price and other relevant factors such as wheelbase, interior area, horsepower, vehicle body type, and its manufacturer. This will aid us to predict market prices for certain types of vehicles informing purchasing decisions from an automobile distributor to redistribute to Australian consumers to maximise profit.

**Summary Statistics of all Variables and Descriptions of Variable Means**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Statistic** | **Number Observations** | **Mean** | **Standard Deviation** | **Median** | **Min** | **Max** |
| Price (in $1000s) | 750 | 46.631 | 5.435 | 46.579 | 28.032 | 62.931 |
| Wheelbase (inches) | 750 | 117.741 | 5.119 | 117.400 | 103.500 | 134.200 |
| Interior Area (Cubic Feet) | 750 | 106.101 | 12.781 | 105.500 | 74.500 | 148.800 |
| Horsepower | 750 | 175.537 | 27.015 | 177 | 89 | 250 |
| Is Car | 750 | 0.601 | 0.490 | 1 | 0 | 1 |
| Is SUV | 750 | 0.399 | 0.490 | 0 | 0 | 1 |
| Is Toyota | 750 | 0.517 | 0.500 | 1 | 0 | 1 |
| Is Honda | 750 | 0.483 | 0.500 | 0 | 0 | 1 |

*Table 1.1 Summary Statistics for all variables in the dataset*

From Table 1.1, it can be observed that the typical automobile in this dataset is priced at $46,631, with a wheelbase of 117.741 inches, an interior area of 106.101 cubic feet, and an average horsepower of 175.537 hp. Of the 750 Toyotas and Hondas sold recently in the given dataset, 60.1% of these vehicles were cars (non-SUVs) and the remaining 39.9% were SUVs. Additionally, 51.7% of the automobiles sold were Toyotas while the remaining 48.3% were Hondas.

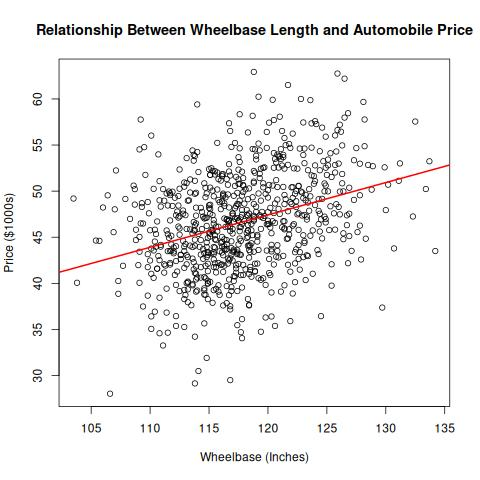
***Note for Remaining Document:*** *‘Is SUV’ is a dummy variable in this dataset, representing whether a vehicle is a SUV (1) or not (0). ‘Is Car’ is another dummy variable and directly corresponds to the complement of ‘Is SUV’, such that it’s equal to 1 if a vehicle is not a SUV and 0 if it is a SUV. Likewise, ‘Is Toyota’ is a dummy variable representing whether (1) or not (0) a vehicle is a Toyota. Since only Toyotas and Hondas make up this dataset, ‘Is Honda’ is another dummy variable that is effectively the complement of ‘Is Toyota’, representing whether a vehicle is a Honda or not, since Toyota and Honda are mutually exclusive.*

A graph of a car price comparison

AI-generated content may be incorrect.**Scatter Plots Displaying the Relationships between Pairs of Variables**

*Figure 2.2 - Relationship between whether the vehicle is a SUV (Independent variable) with Price (Dependent Variable)*

A graph of a car with a red line

AI-generated content may be incorrect.

*Figure 2.1 - Relationship between Wheelbase (Independent variable) with Price (Dependent Variable)*

*Figure 2.3 - Relationship between whether a vehicle is a SUV or not (Independent variable) with the Wheelbase (Dependent Variable)*

**Perfect Multicollinearity with Including all Variables in Linear Regression**

If we run a regression with price as the dependent variable and all other variables as regressors, we will encounter the dummy variable trap, an instance of perfect multicollinearity. This occurs since each vehicle in our dataset must be either a Toyota or a Honda, but not both. Hence, ‘Is Toyota’ and ‘Is Honda' are mutually exclusive. The two dummy variables are perfectly collinear with the constant regressor because ‘Is Toyota’ and ‘Is Honda’ are a linear combination of the constant regressor (see Equation 3.1).

(Equation 3.1)

*Equation 3.1 - Equation representing perfect multicollinearity when including ‘Is Toyota’, ‘Is Honda' and the constant term in the regression.*

An identical problem is also encountered if we include both ‘Is Car’ and ‘Is SUV’ variables, as each observation must be in one category, and they are also mutually exclusive (Equation 3.2).

(Equation 3.2)

*Equation 3.2 - Equation representing perfect multicollinearity when including ‘Is Car’, ‘Is SUV’ and the constant term in the regression.*

To avoid this issue, we can omit one dummy variable from each of the equations, such as omitting from the first equation and omitting from the second equation. In doing so, the effects of the omitted variables would be included in the constant term.