# 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.*

## Scatter Plots Displaying the Relationships between Pairs of Variables

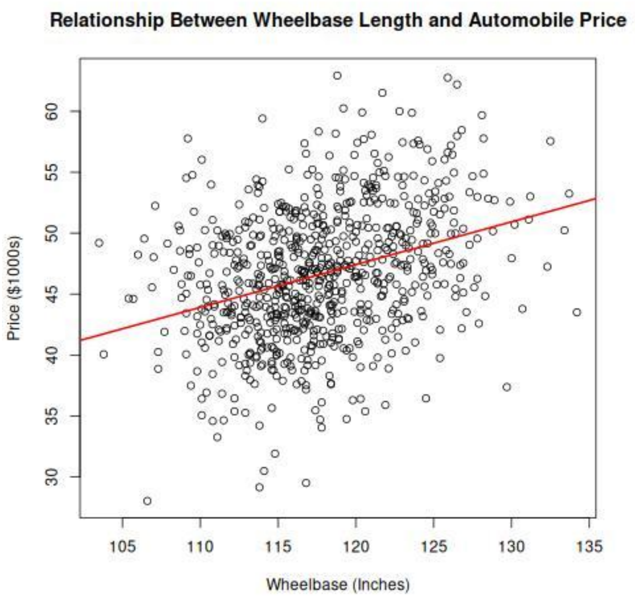


*Figure 2.1 - Relationship between Wheelbase (Independent variable) with Price (Dependent*

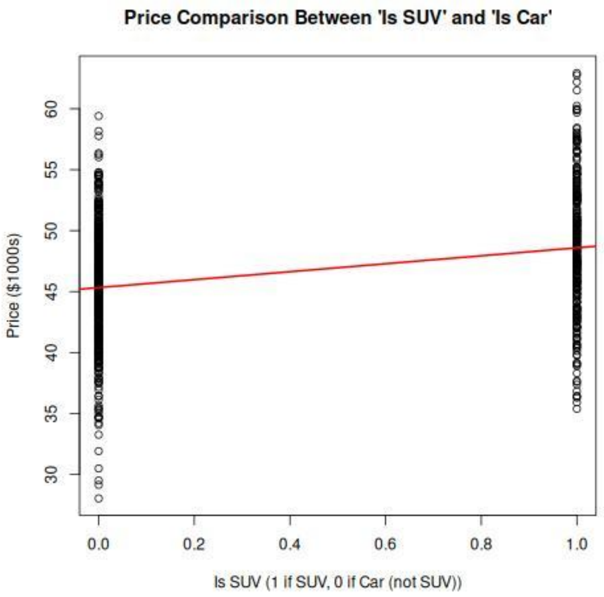


*Figure 2.2 - Relationship between whether the vehicle is a SUV (Independent variable) with Price (Dependent Variable)*

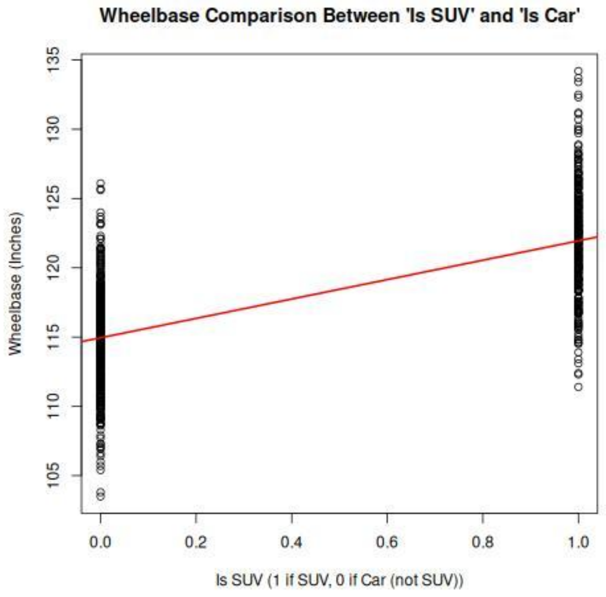


*Figure 2.3 - Relationship between whether a vehicle is a SUV or not (Independent variable) with the Wheelbase (Dependent Variable)*

## 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).

$$Is\,Toyota_i + Is\,Honda_i = 1 = X_{0i} \text{ (Equation 3.1)}$$

*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).

$$Is\,Car_i + Is\,SUV_i = 1 = X_{0i} \text{ (Equation 3.2)}$$

*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 $Is\,Honda_i$ from the first equation and omitting $Is\,SUV_i$ from the second equation. In doing so, the effects of the omitted variables would be included in the constant term.

## Various Linear Regressions for Price with Different Independent Variables per Regression

Regression 1: Price vs Wheelbase
Regression 2: Price vs Wheelbase + Interior Area
Regression 3: Price vs Wheelbase + Interior Area + Horsepower
Regression 4: Price vs Wheelbase + Interior Area + Horsepower + Is SUV
Regression 5: Price vs Wheelbase + Interior Area + Horsepower + Is SUV + Is Toyota

*NB: In Table 4.1, the numbers in parentheses below coefficients are the standard errors (Accounting for heteroskedasticity) corresponding to their respective estimates. The number of asterisks beside the coefficient estimates refers to the extent of statistical significance for the hypothesis test. (* means statistically significant at 10% statistical significance, ** means statistically significant at 5% statistical significance and *** means statistically significant at 1% statistical significance).*

|  | Reg1 | Reg2 | Reg3 | Reg4 | Reg5 |
|---|---|---|---|---|---|
| Wheelbase (Inches) | 0.351***<br>(0.038) | 0.065<br>(0.055) | 0.089*<br>(0.048) | 0.044<br>(0.054) | 0.046<br>(0.054) |
| Interior Area (Cubic Feet) | N/A | 0.152***<br>(0.022) | 0.142***<br>(0.020) | 0.139***<br>(0.020) | 0.139***<br>(0.020) |
| Horsepower | N/A | N/A | 0.087***<br>(0.006) | 0.086***<br>(0.006) | 0.094***<br>(0.007) |
| Is SUV | N/A | N/A | N/A | 0.816*<br>(0.441) | 0.801*<br>(0.440) |
| Is Toyota | N/A | N/A | N/A | N/A | -0.754**<br>(0.379) |
| Constant | 5.310<br>(4.467) | 22.798***<br>(4.989) | 5.804<br>(4.453) | 11.296**<br>(5.330) | 10.008*<br>(5.304) |
| Adjusted $R^2$ | 0.108 | 0.163 | 0.349 | 0.351 | 0.354 |
| Number of Observations | 750 | 750 | 750 | 750 | 750 |

*Table 4.1 - Regression output for the 5 regressions listed above*

## Interpretation of Constant Term in Regressions and Omitted Variable Bias

a) Interpreting the Constant Term in Regression 4 vs Constant Term in Regression 5

The constant term in regression 4 is 11.296, which decreases to 10.008 in regression 5, once we added the dummy variable 'Is Toyota' to the regression for predicted price.

This suggests that the price contribution by cars (non-SUV) for an automobile in regression 4 is included within the price increase of $11,296 (11.296 x $1000) by the constant term. Since we omitted the dummy variable 'Is Car' from the regression in favour of the dummy variable 'Is SUV', the intercept constitutes the price contribution of cars (non-SUVs). The predicted value of a car (non-SUV) in regression 4 is given by the expression:

$$(11.296 + 0.044 \times wheelbase + 0.139 \times interior\ area + 0.086 \times horsepower) \times 1000$$

Note that the price of an SUV with exactly the same specifications (wheelbase, interior area and horsepower) would be predicted to be valued at 0.816 (estimated coefficient of 'Is SUV' in regression 4) x $1000 = $816 more than cars. This is because the coefficient of 'Is SUV' is relative to the omitted variable 'Is Car'.

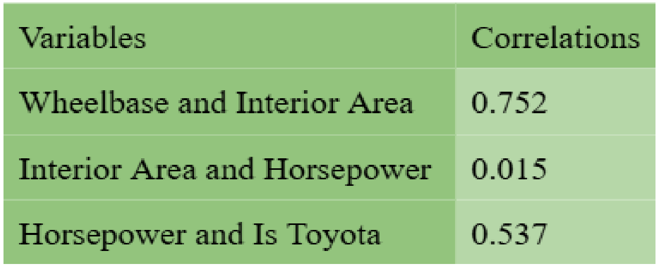
Similarly, the constant term in regression 5 is 10.008, suggesting that the price contribution for Honda cars is included within the $10,008. The predicted value of a Honda car is given by the expression:

$$(10.008 + 0.046 \times wheelbase + 0.139 \times interior\ area + 0.094 \times horsepower) \times 1000$$

The estimated market price of a Toyota car would be $754 less than a Honda car, since the estimated coefficient for Toyota in the regression is -0.754 (holding other regressors constant). With similar reasoning, the estimated market price for a Honda SUV would be $801 more than a Honda car, and the estimated market price for a Toyota SUV would be $45 ($801 - $754) more than a Honda car, again assuming all other regressors are held constant.

Therefore, we can infer that the constant in regression 4 represents the price contribution of an average ‘car’ in this dataset specifically, whilst the constant in regression 5 represents the price contribution of an average ‘Honda car’ with the inclusion of the Toyota dummy regressor.

b) Three Variables Affected by Omitted Variable Bias from Regressions (1-4) to 5



| Variables | Correlations |
|---|---|
| Wheelbase and Interior Area | 0.752 |
| Interior Area and Horsepower | 0.015 |
| Horsepower and Is Toyota | 0.537 |

*Table 5.1 - Correlations between variables to show omitted variable bias.*

‘Wheelbase’ may have been impacted by omitted variable bias. In regression 1, it had coefficient 0.351 (Table 4.1). However, it dropped to 0.046 (Table 4.1) in regression 5. It can be inferred that the coefficient of ‘wheelbase’ is positively biased. The most noticeable reduction in magnitude was observed to be the addition of the interior area regressor in regression 2, dropping from 0.351 to 0.065 (Table 4.1). To confirm this, ‘interior area’ has a positive effect on ‘price’ (positive coefficient) and has a positive correlation of 0.752 with ‘wheelbase’ (Table 5.1). Therefore, confirming the positive bias. Additionally, the sign of the coefficient tied to ‘wheelbase’ was unchanged, meaning that the magnitude of the bias term was not large enough to change its sign.

Similarly, the coefficient of ‘interior area’ was partially positively biased. From Table 4.1, it had the value of 0.152 in regression two, dropping in magnitude to 0.139 in regression 5, mainly due to the addition of the ‘horsepower’ regressor. From Table 5.1, the correlation between ‘horsepower’ and ‘interior area’ is positive, and the effect of ‘horsepower’ on the predicted price of the vehicle is positive, as it had a coefficient of 0.094 (Table 4.1) in regression 5. Therefore, this confirms that the coefficient corresponding to ‘interior area’ was slightly positively biased.

‘Horsepower’ may have been negatively biased. From Table 4.1, it had a coefficient of 0.087 in regression 3, increasing in magnitude to 0.094 when ‘Is Toyota’ was added in regression 5. The increase was relatively large relative to the standard error of 0.007. To confirm this, the correlation between ‘Is Toyota’ and ‘horsepower’ was positive with a magnitude of 0.537 (Table 5.1). In addition to this, ‘Is Toyota’ had a negative effect on the predicted price of vehicles as it had a coefficient of -0.754 (Table 4.1). Therefore, the sign of the bias term is negative in this case, thus confirming that the coefficient of ‘horsepower’ was underestimated.

## Calculate Predicted Price and Maximise Profit by Purchasing Decisions

a) Calculate Predicted Price for all Listed Automobiles based on Regression

| (Intercept) | Wheelbase (Inches) | Interior Area (Cubic Feet) | Horsepower | SUV | Toyota |
|---|---|---|---|---|---|
| 10.00773 | 0.04623 | 0.13909 | 0.09396 | 0.80138 | -0.75441 |

*Table 7.1 - Coefficient Estimates of Regression 5 to 5 decimal places*

$$Predicted\ Price\ (ID1) = (10.00773 + 0.04623 \times 110 + 0.13909 \times 85 + 0.09396 \times 112 + 0.80138 \times 0 - 0.75541 \times 1) \times \$1000 = \$36{,}683.79$$

$$Predicted\ Price\ (ID2) = (10.00773 + 0.04623 \times 134 + 0.13909 \times 120 + 0.09396 \times 200 + 0.80138 \times 1 - 0.75541 \times 0) \times \$1000 = \$52{,}486.73$$

$$Predicted\ Price\ (ID3) = (10.00773 + 0.04623 \times 110 + 0.13909 \times 105 + 0.09396 \times 130 + 0.80138 \times 0 - 0.75541 \times 1) \times \$1000 = \$41{,}156.87$$

$$Predicted\ Price\ (ID4) = (10.00773 + 0.04623 \times 125 + 0.13909 \times 110 + 0.09396 \times 160 + 0.80138 \times 1 - 0.75541 \times 1) \times \$1000 = \$46{,}165.95$$

These predicted prices are the conditional expectations of Price given specific values of the regressors, since it is assumed that $E(u_i|X_i) = 0$ (as per Least Squares Assumption 1)

| ID | Is SUV | Is Toyota | Interior Area | Horse power | Wheel base | Cost | Number Available | Predicted Price | Number Ordered |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 85 | 112 | 110 | 23000 | 3 | $36,683.79 | 1 |
| 2 | 1 | 0 | 120 | 200 | 134 | 40000 | 2 | $52,486.73 | 0 |
| 3 | 0 | 1 | 105 | 130 | 110 | 27000 | 4 | $41,156.87 | 2 |
| 4 | 1 | 1 | 110 | 160 | 125 | 30000 | 4 | $46,165.95 | 4 |

*Table 7.2 - Car purchase data with the predicted prices and numbers ordered for each ID*

b) Maximise profit by purchasing vehicles with maximum budget of $200,000

$$Predicted\ ID1\ profit = Predicted\ Price\ (ID1) - Cost(ID1) = \$13{,}683.79$$
$$Predicted\ ID2\ profit = Predicted\ Price\ (ID2) - Cost(ID2) = \$12{,}486.73$$
$$Predicted\ ID3\ profit = Predicted\ Price\ (ID3) - Cost(ID3) = \$14{,}156.87$$
$$Predicted\ ID4\ profit = Predicted\ Price\ (ID4) - Cost(ID4) = \$16{,}165.95$$

To maximise profit, all available vehicles from the vehicle ID with the greatest expected profit should be bought. In this scenario, ID4 has the greatest expected profit of $16,165.95, so we should buy all 4 available vehicles, which leaves us with a remaining budget of $80,000. From here, all available vehicles from vehicle ID3 should be bought such that we do not exceed our budget, as it has the second greatest expected profit. We can only buy 2 such vehicles since $3 \times \$27,000 = \$81,000,$ which is greater than our remaining budget. After purchasing 2 ID3 vehicles, we are left with a budget of $26,000, which can be used to purchase one ID1 vehicle costing $23,000, yielding an expected profit of $13,683.79.

Expected Profit = $\$16165.95 \times 4 + \$14156.87 \times 2 + \$13683.79 = \$106,661.33$
Total Cost = $\$30,000 \times 4 + \$27,000 \times 2 + \$23,000 = \$197,000$

To validate that this indeed was the maximum profit, we simulated it using Python. The implementation involved utilising for loops for each vehicle ID ranging from 0 to the number of vehicles available, calculating profit and cost, and including the profit in a list data structure, provided that the combination of vehicles suggested was within budget. Lastly, we sorted the data structure in descending order based on profit and yielded the combination that maximises profit.