IB Mathematical Studies 2008– Internal Assessment

What is the relationship between the weight of a car, its CO_2 emissions and its fuel consumption?

1

Contents

Statement of Intent:

My aim is to discover whether the weight of any particular car determines its environmental impact measured by CO_2 output and fuel consumption. I will also explore the relationship between CO_2 output and fuel consumption. I am interested in this because I will soon have my licence and will need to decide which car will be most fuel efficient and environmentally friendly.

In order to collect my data and control the variables, I will be only collecting data on cars that are 2 wheel drives with 2.0L petrol engines and automatic transmission. I will be collecting data on the fuel consumption (L/100km) and CO_2 emissions (gm/km) as well as the curb weight (kg) of the cars (curb weight is the total weight of a vehicle with standard equipment, all necessary operating consumables such as motor oil and coolant, a full tank of fuel, and not loaded with either passengers or cargo). I will be obtaining this information primarily from the internet, car brochures and calling up car dealerships.

Firstly, to discover whether there is a relationship between each of the variables (fuel consumption and weight; CO_2 emissions and weight; and finally CO_2 emissions and fuel consumption). I will be plotting scatter-plots, and finding the correlation coefficient and a regression equation if appropriate. I will use this regression equation or 'line of best fit' to make some predictions and compare them to already recorded values and calculate the percentage error. This will tell me how reliable the regression equation is in predicting data. I will then test the null hypothesis, that the two variables are independent, for all the variables using the chi-squared test. To work out the groups to use in my hypothesis testing, I will use uni-variate statistics to find the averages of the different groups of data. I will then divide my groups into 'below average' and 'above average'.

Raw Data

	Combined Fuel Consumption (L/100km)	*CO_2 emissions (g/km)*	*Weight (kg)*
Audi TT Roadster 2.0 TFSI S-Tronic	9.4	188	1345
Audi A4 B7 2.0 Tfsi Quattro	8.3	226	1535
Audi A4 B7 2.0 Tfsi Exclusive	9.7	194	1450
BMW 120i E87 hatch	7.9	190	1300
Citroën C5 Saloon	8.6	206	1958
Citroën C4 Exclusive	8.1	193	1292
Citroën C4 Picasso	8.9	211	1600
Ford Focus	8.0	189	1339
Honda Civic Sport Sedan	8.4	200	1290
Hyundai Elantra	7.8	186	1285
Hyundai Tucson City	9.9	236	1698

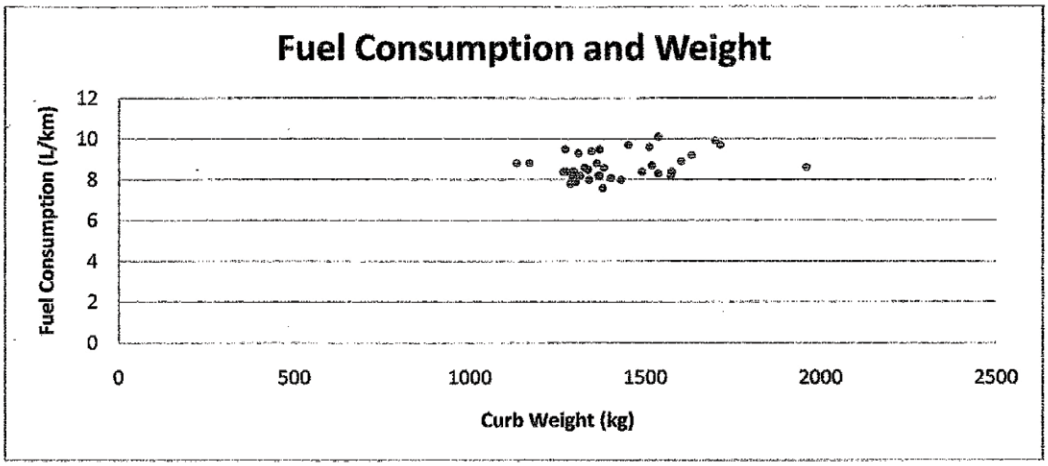
Hyundai i30	7.6	182	1377
Kia Cerato Sedan	8.2	195	1313
Kia SUV Sportage LX	9.2	220	1630
Kia Cerato Hatch	8.2	195	1365
Mazda MX - 5 Roadster Coupe	8.8	207	1169
Mazda 3	8.4	199	1275
Mazda MX - 5 Soft Top	8.8	207	1,132
Peugeot 307xse Sedan	8.2	195	1368
Peugeot 307CC	8.4	199	1488
Peugeot 307CC	8.4	199	1573
Renault Megane Hatch - Expression	8.4	201	1265
Renault Megane Sedan	8.4	201	1295
Saab 9-3 Vector 2.0T Convertible	9.7	232	1712
Saab 9-3 Sport Sedan Linear	9.6	229	1509
Skoda Octavia RS	8.1	193	1400
Skoda Octavia	8.5	203	1335
Subaru Impreza R	8.8	208	1360
Suzuki SX4	9.5	221	1270
Suzuki Grand Vitara 4x4	10.1	242	1535
Volkswagen Jetta Turbo	8.0	192	1430
VOLKSWAGEN GOLF 2.0 Petrol	8.6	206	1326
Volkswagen Eos	8.2	194	1571
Volkswagen Beetle Cabrio	9.5	228	1368
Volkswagen Jetta	8.6	206	1380
Volkswagen Passat Fsi	8.7	207	1,517
Volkswagen New Beetle Cabrio Cabriolet	9.3	223	1,308

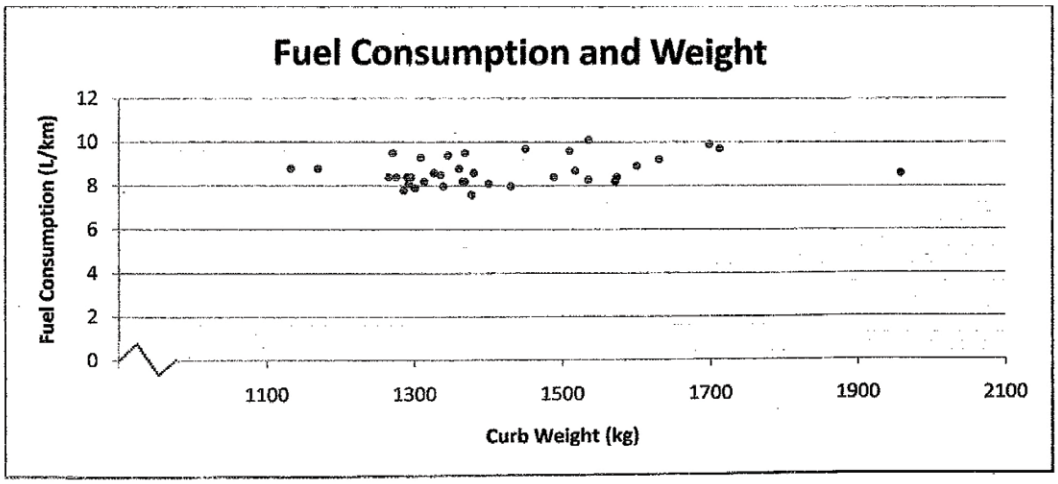
4

1) FUEL CONSUMPTION AND WEIGHT

Scatter Plots:

The scatter plot will allow me to visually see if there is a trend in the results; it will also allow the plotting of a line of best fit if appropriate. I have also done a second graph with a different horizontal scale to better show the spread of the data.





Visually, there does not seem to be a very convincing trend. In the graph with the smaller horizontal scale, you can see a very weak positive correlation. However, the strength of the relationship will be more accurately explored by the regression equation and the hypothesis testing.

5

Finding the correlation coefficient and regression equation:

As my intent is to find the strength of the relationship between fuel consumption, CO_2 emissions and curb weight, I will find the correlation coefficient. This will give an indication of whether there is a strong linear relationship between two factors. If it looks like there is, I will find a 'line of best fit' from which I can predict results.

Given that I do not have the Sxy value, I will use the correlation coefficient formula, allowing that *x is weight* and *y is fuel consumption*: $\frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}}$

Using the graphics calculator, 2 Var Stats, I calculated that...

$\bar{y} = 8.68$; $\bar{x} = 1415.22$; $\sum y^2 = 2803.14$; $\sum x^2 = 75091377$; $\sum xy = 455726.2$; and n = 37

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}}$$

$$r = \frac{455726.2 - 37 \times 1415.21 \times 8.68}{\sqrt{75091377 - 37 \times 1415^2} \times \sqrt{2803.14 - 37 \times 8.68^2}}$$

$$r = \frac{1217.36}{1004.51 \times 3.933}$$

$$r = 0.3081$$

$$r^2 = .0949$$

After doing the same linear regression on my calculator, using the same data, the *r* value equals 0.3035. Although this value is similar to the one I gained from the equation above, it could differ slightly because of the rounding I had to do in the equation. For this reason I will use the calculator value. It is the most accurate because it works out the regression value without any rounding. T his means that only 9.5 % of the variation in the dependent variable, fuel consumption, can be explained by the variation in the independent variable, weight. This is an extremely weak positive correlation therefore a regression equation will not be appropriate.

Hypothesis Testing:

To work out the groups of values for my different chi squared tests, I will find the average of each of these two factors; fuel consumption and car weight. I calculated the median and mode of both weight and fuel consumption. The median of the fuel consumption data was 8.5, and the mode was 8.4. The median of the weight data was 1368 as was the mode. However, both did not ensure at least five pieces of data in each of the observed frequency cells, as the value for the average does. Before I worked out averages, I calculated whether there were any outliers, for each of these factors. Outliers are values outside the limits of Q_1-1.5 x IQR or Q_3+1.5 x IQR. After working out these equations, I calculated these limits.

Fuel consumption: Q_1=8.2, Q_3=9.25, IQR=1.05	***Weight***: Q_1= 1297.5, Q_3= 1526, IQR=228.5
Upper limit: 9.25+1.5x1.05=10.825	Upper limit: 1526+1.5x228.5=1868.75
Lower limit: 8.2-1.5x1.05=6.625	Lower limit: 1297.5-1.5x228.5=954.75

This shows, after looking at the maximum and minimum values of the data, that there is one outlier over the upper weight limit. However, the relevant fuel consumption for that car is not an outlier so I will not remove that piece of data from my calculations. Additionally, when looking at the graph, the outlying value does not look too out of place with the rest of my data. For bivariate data, an outlier is more easily determined visually. Therefore, I will use the average of the raw data to find an exact value, although I realise weight, fuel consumption and CO_2 emissions are all continuous data and could be placed in frequency tables with class intervals. However, I am only interested in this sort of statistical analysis to organise my chi squared tests, so will not be doing this.

Weight: The combined value of all the car weights is 52 363kg, and there are 37 different pieces of data. Therefore:

$$\frac{52363}{37} = 1415.22 \text{ kg}$$

The average is 1415.22 to two decimal places, so I will divide my two groups up into:

-Below average weight = (<1415.22)

-Above average weight = (>1415.22)

Fuel: The combined value of all the fuel consumption is 321.2 (L / 100km), and there are 37 different pieces of data. Therefore:

$$\frac{321.2}{37} = 8.68 \text{ (L/100km)}$$

The average is 8.68, so I will divide my two groups up into:

-Below average fuel consumption = (<8.68)

-Above average fuel consumption = (>8.68)

So therefore, my hypotheses for fuel consumption and weight are...

- ***H_o (null hypothesis)*** = fuel consumption is independent of car weight
- ***H_1 (alternate hypothesis)*** = fuel consumption is dependent on car weight

So, if the χ^2calc is bigger than the χ^2critical, we discard the null hypothesis and accept the alternate hypothesis.

Observed:

	Below Average Weight	Above Average Weight	
Below Average Fuel	16	6	**22**
Above Average Fuel	7	8	**15**
	23	**14**	**37**

Expected:

	Below Average Weight	Above Average Weight	
Below Average Fuel	$\frac{22\times23}{37}$	$\frac{22\times14}{37}$	**22**
Above Average Fuel	$\frac{15\times23}{37}$	$\frac{15\times14}{37}$	**15**
	23	**14**	**37**

=

	Below Average Weight	Above Average Weight	
Below Average Fuel	13.68	8.32	**22**
Above Average Fuel	9.32	5.68	**15**
	23	**14**	**37**

χ^2 Calculations:

fo	fe	$fo-fe$	$(fo-fe)^2$	$\frac{(fo-fe)^2}{fe}$
16	13.67	2.33	5.4289	0.3971
6	8.32	-2.32	5.3824	0.647
7	9.32	-2.32	5.3824	0.578
8	5.67	2.33	5.4289	0.957
				Total : 2.58 (3sf)

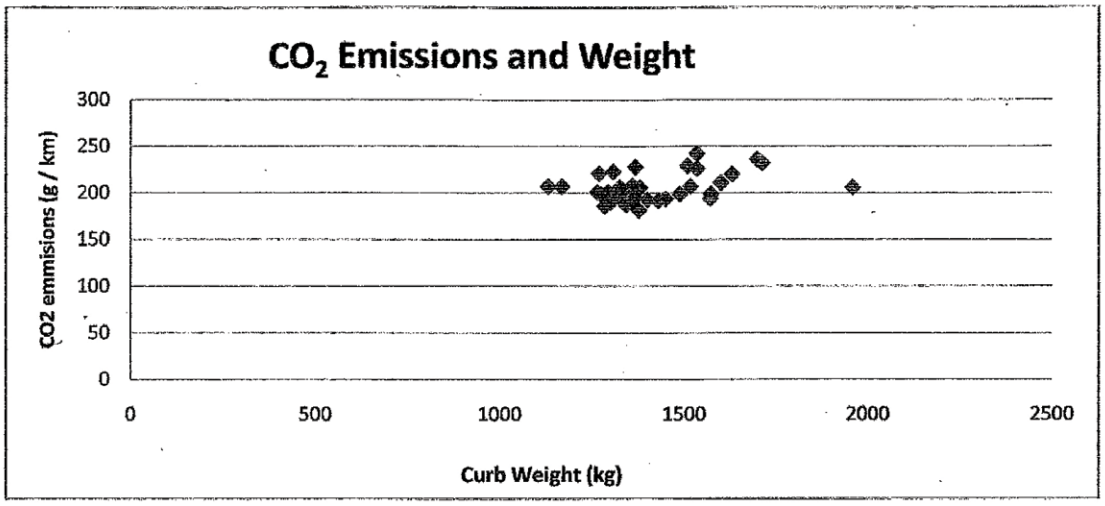
To find the degree of freedom (df), we use the equation $v = (r-1)(c-1)$, where r=number of rows and c= number of columns $v = (2-1)(2-1)$

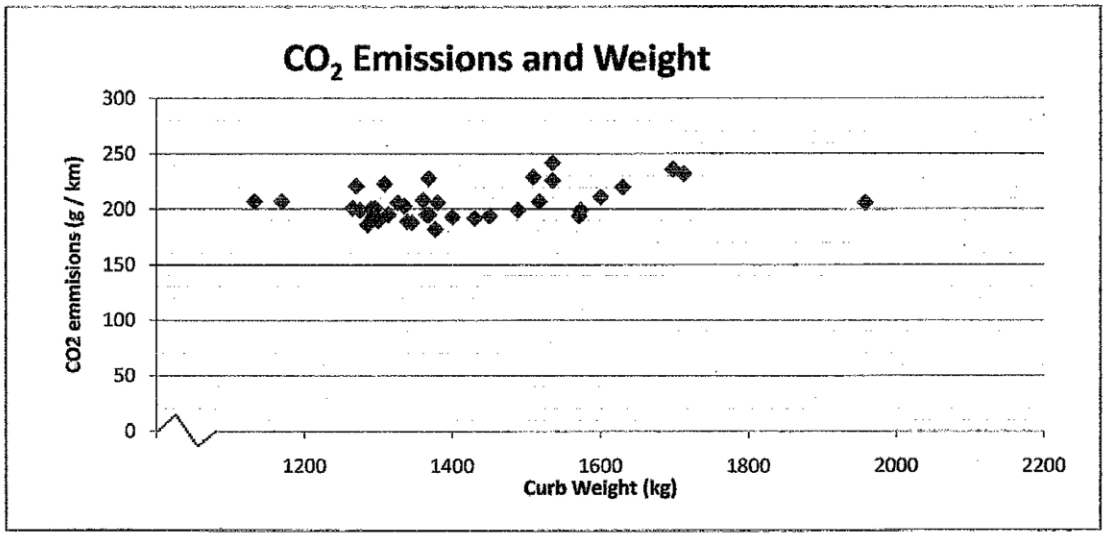
Therefore the degree of freedom is 1. Testing this at a 10% significance level, the χ^2critical value is 2.71, and χ^2calc is not bigger than χ^2critical. This means that we reject the alternate hypothesis and accept the null hypothesis that car weight and fuel consumption are independent. I expected this, as the correlation coefficient that I found for this data was extremely low, showing a very weak relationship. Additionally, visually there was no trend in the scatter plot graph. From the outcomes of the correlation coefficient and the hypothesis testing, as well as the visual scatter plot, it is unlikely there is any relationship between fuel consumption and the weight of a car.

8

2) CO_2 EMISSIONS AND WEIGHT

Scatter Plots:





Once again, visually, there does not seem to be a very convincing trend. In the graph with the smaller horizontal scale, you can see a very weak positive correlation. However, the strength of the relationship will be more accurately discussed in the regression equation and the chi squared test.

Finding the correlation coefficient and regression equation:

Again, given that I do not have the Sxy value, I will use the regression formula, given that *x is weight,* and *y is CO_2 emissions*: $\frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}}$

9

Using the graphics calculator, 2 Var Stats, I calculated that...

$$\bar{x} = 1415.21;\ \bar{y} = 205.49; \sum x^2 = 75091377; \sum y^2 = 1570603;\ \sum xy = 10794445;\ n = 37$$

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}}$$

$$r = \frac{10794445 - 37 \times 1415.21 \times 205.49}{\sqrt{\sum 75091377 - 37 \times 1415.21^2}\sqrt{\sum 1570603 - 37 \times 205.49^2}}$$

$$r = \frac{34419.39}{\sqrt{987061.27} \times 90.75}$$

$$r = 0.384$$

$$r^2 = 0.147$$

After comparing this with the value obtained from the calculator, I found they are almost exactly the same, the calculator *r* value being 0.382. Therefore this also means that only that only 14.8 % of the variation in the dependent variable, CO_2 emissions, can be explained by the variation in the independent variable, weight. This is also an extremely weak positive correlation therefore a regression equation will not be appropriate.

Hypothesis Testing:

Again, to work out the groups of values for my different chi squared tests, I will find the average of the remaining factor, carbon dioxide emissions. Also, I will first ensure there are no outliers in the CO_2 data. Outliers are values outside the limits of Q_1-1.5 x IQR or Q_3+1.5 x IQR. After working out these equations, I calculated these limits.

CO_2 emissions: Q_1=194, Q_3=215.5, IQR=21.5

Upper limit: 215.5+1.5x21.5=247.75

Lower limit: 194-1.5x21.5=161.75

This shows, after looking at the maximum and minimum values of the data, that there are no outliers.

The combined value of all the CO_2 emissions is 7603 (g / 100km), and there are 37 different pieces of data. Therefore: $\frac{7603}{37} = 205.48$ (g/100km)

The average is 205.48, so I will divide my two groups up into:

-Below average CO_2 emissions = (<205.48)
-Above average CO_2 emissions = (>205.48)

So therefore, my hypotheses for CO_2 emissions and car weight are...

- ***H_0 (null hypothesis)*** = CO_2 emissions are independent of car weight
- ***H_1 (alternate hypothesis)*** = CO_2 emissions are dependent on car weight

So, if the χ^2calc is bigger than the χ^2critical, we discard the null hypothesis and accept the alternate hypothesis.

10

Observed:

	Below Average Weight	Above Average Weight	
Below Average CO_2	15	5	**20**
Above Average CO_2	8	9	**17**
	23	**14**	**37**

Expected:

	Below Average Weight	Above Average Weight	
Below Average CO_2	$\frac{23\times 20}{37}$	$\frac{20\times 14}{37}$	**20**
Above Average CO_2	$\frac{23\times 17}{37}$	$\frac{17\times 14}{37}$	**17**
	23	**14**	**37**

=

	Below Average Weight	Above Average Weight	
Below Average CO_2	12.43	7.57	**20**
Above Average CO_2	10.57	6.43	**17**
	23	**14**	**37**

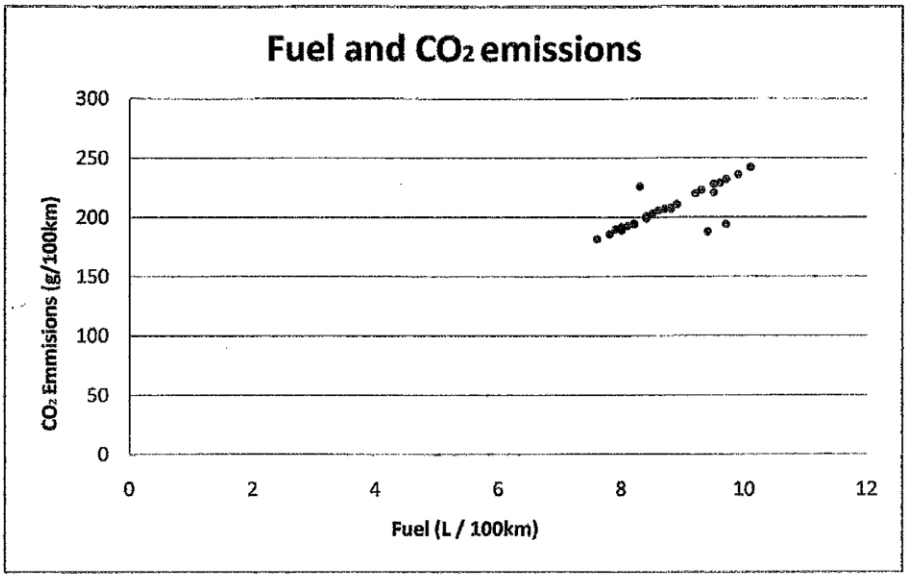
X^2 Calculations

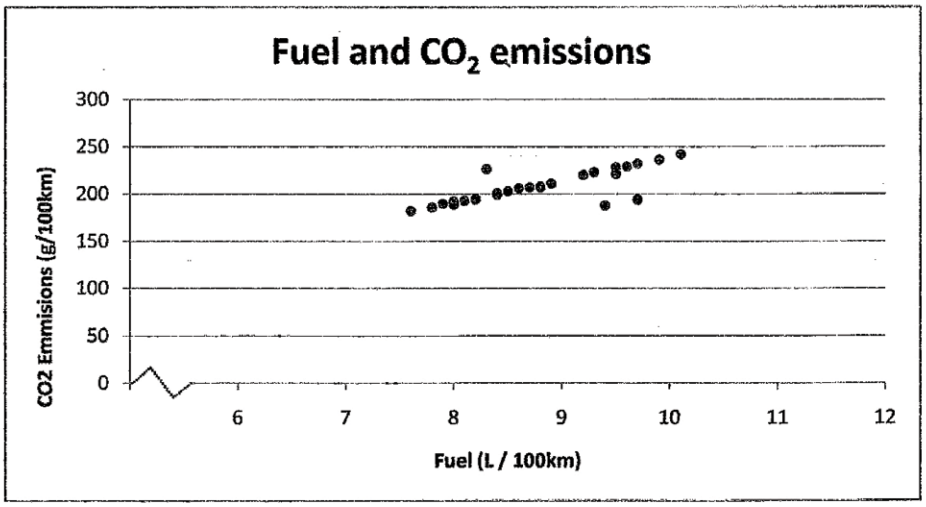
fo	fe	$fo-fe$	$(fo-fe)^2$	$\frac{(fo-fe)^2}{fe}$
15	12.43	2.57	6.605	0.531
5	7.56	-2.56	6.55	0.867
8	10.56	-2.56	6.55	0.6206
9	6.43	2.57	6.605	1.027
				Total : 3.046

To find the degree of freedom (df), we again use the equation $v = (r-1)(c-1)$, where r=number of rows and c= number of columns, so therefore $v = 1$. Testing this at a 10% significance level, the χ^2critical value is 2.71, and χ^2calc is 3.046, and so it *is* bigger than χ^2critical. This means that we reject the null hypothesis and accept the alternate hypothesis - that the amount of CO_2 emissions do depend on the weight of the car. I was not expecting this as this data also had a low regression value. However, if we test the χ^2calc at a 5% significance level, the χ^2critical value is 3.84, and therefore is bigger than the χ^2calc. Therefore, we can accept the null hypothesis at only a 10% significance level. This proves that although there is some relationship, it is not very strong.

3) CO_2 EMISSIONS AND FUEL CONSUMPTION

Scatter Plots





Visually, there seems to be a strong positive linear trend. However, the strength of the relationship will be more accurately tested by finding the regression equation.

12

Finding the correlation coefficient and regression equation:

Again, given that I do not have the Sxy value, I will use the regression formula given *that x is fuel consumption* and *y is CO_2 emissions*: $\frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}}$

Using the graphics calculator, 2 Var Stats, I calculated that...

$\bar{x} = 8.68;\ \bar{y} = 205.49; \sum x^2 = 2803.14; \sum y^2 = 1570603;\ \sum xy = 66279.3\ ; n = 37$

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}}$$

$$r = \frac{66279.3 - 37 \times 8.68 \times 205.49}{\sqrt{2803.14 - 37 \times 8.68^2}\sqrt{1570603 - 37 \times 205.49^2}}$$

$$r = \frac{284.13}{3.933 \times 90.75}$$

$$r = 0.796$$

$$r^2 = 0.634$$

From my calculator, the *r* value was 0.792 and the r^2 was 0.6266. I will use these values for working out my linear equation as these are more accurate as they are calculated without rounding values. As this correlation coefficient shows a medium positive correlation, I will find a regression equation to find a line of best fit. This will allow me to predict further results. Using the previous values above and including $S_x = 0.6319$, $S_y = 14.967$, I will use this least squares regression equation to find the line of best fit:

$$y - \bar{y} = r \times \frac{Sy}{Sx} \times (x - \bar{x})$$

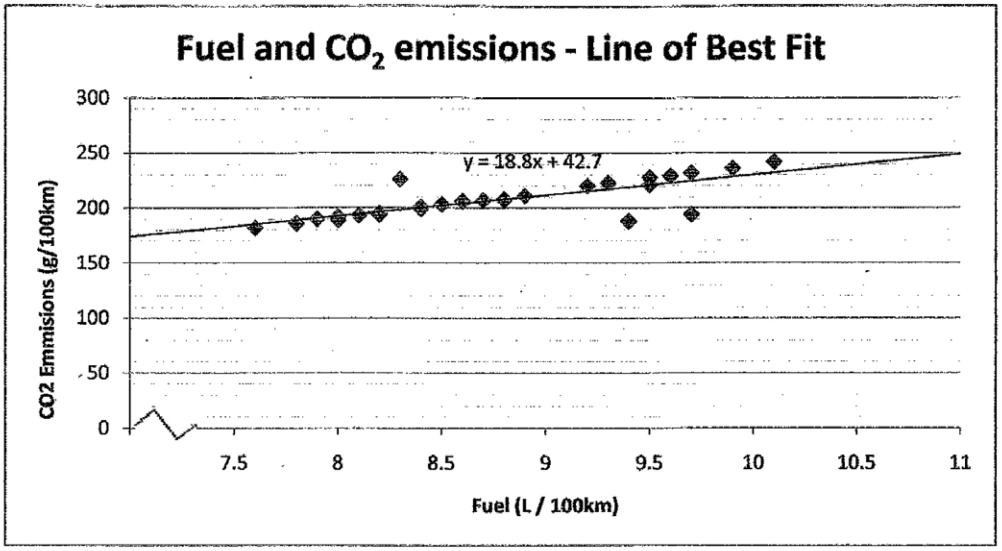
$$y - 205.49 = 0.792 \times 14.967/0.6319 \times (x - 8.68)$$

$$y - 205.49 = 18.759 \times (x - 8.68)$$

$$y - 205.49 = 18.759x - 162.828$$

$$y = 18.8x + 42.7$$

13



I will be extrapolating and interpolating predictions from my formula for the line of best fit. I will also be comparing the percentage error between the extrapolated and interpolated values and actual values.

First I will be comparing a y value found from an interpolated value on the line of best fit, to its actually recorded value. The Skoda Octavia has recorded as its fuel consumption 8.5L/100km. This is the 'x' value. It also has CO_2 emissions of 203 gm /100km as its recorded '*y*' value. So, using my formula and substituting in 8.5 as my '*x*' value.

$$y = 18.8x + 42.7$$

$$y = 18.8 \times 8.5 + 42.7$$

$$y = 202.5$$

This value is extremely close to the actual '*y*' value of 203; however, it is not exactly the same. Therefore, I will now find the percentage error between them.

$$error = estimated\ value - true\ value$$

$$error = 202.5 - 203$$

$$error = -0.5$$

Therefore substitute this value into the absolute percentage error equation;

$$percentage\ error = \frac{|error|}{actual\ value} \times 100\%$$

$$percentage\ error = \frac{|0.5|}{203} \times 100\%$$

$$percentage\ error = 0.246\%$$

14

This shows a very close relationship between the real and predicted value.

The Saab 9-3 Sport Linear has recorded as its fuel consumption 9.6 L/100km. This is the 'x' value. It also has CO_2 emissions of 229 gm /100km as its recorded *y* value. So, using my formula and substituting in 229 as my *y* value.

$$y = 18.8x + 42.7$$

$$229 = 18.8x + 42.7$$

$$229 - 42.7 = 18.8x$$

$$186.3 = 18.8x$$

$$x = 9.91$$

As this value is not the same as my recorded value for the x value, I will find the percentage error for the estimated (9.91) and true (9.6) values. I will do this using the formula to find percentage error.

$$error = 9.91 - 9.6$$

$$error = .31$$

Therefore sub this value into the absolute percentage error equation;

$$percentage\ error = \frac{|error|}{actual\ value} \times 100\%$$

$$percentage\ error = \frac{|.31|}{9.6} \times 100\%$$

$$percentage\ error = 3.22\%$$

Both of these percentage errors are quite small, showing that the linear regression equation is quite good and there is a strong correlation between the data and the equation. I believe it is reasonable to extrapolate with data both above and below my range of data. My smallest fuel consumption value is 7.6 L/km, so I will try and predict the CO_2 emissions for the value of 6 L/km. Substituting 6 L/km into the regression equation as my '*x*' value, I gain the CO_2 emission value of 155.5 gm/km. This is also below any of the data I recorded.

My largest value for CO_2 emissions is 242 gm/km, so I will try and predict the fuel consumption with the CO_2 emission value of 290 gm/km. Substituting this value into the regression equation as the '*y*' value; I gain the fuel consumption value of 13.15 L/km. This is also above any of the data values I have collected.

There is not enough data to gain five pieces of CO_2 data in each of the observed frequency cells, so I am unable to do a chi squared test. However, the linear regression and the interpolation have proved that there is a fairly strong correlation between these two factors.

15

Conclusion:

I was quite surprised with the results of my project. There was no linear relationship between weight and fuel consumption and the hypothesis testing proved that they are indeed independent of each other. I also found that there was no linear relationship between CO_2 emissions and weight. However, the hypothesis testing proved that the amount of CO_2 emissions do depend on the weight of the car. This would have differed, however, if I had tested it at 5% as opposed to 10% significance level. Additionally, as there was not enough data to provide five pieces in each observed frequency cell to be able to test CO_2 and fuel consumption, it is unclear whether they depend on each other. Although, they still did have a fairly strong linear relationship. Therefore we can conclude that there is absolutely no relationship between the weight of a car and its fuel consumption. However, we can also conclude that there is a fairly strong linear relationship between the fuel consumption and CO_2 emissions of a car *and* that a car's CO_2 emissions do depend on its weight. So, when planning to purchase my own car, it is obvious that the best way to reduce environmental impact (in the form of CO_2 emissions) is to buy a light car which has low fuel consumption.
The correlation coefficient for fuel consumption and CO_2 emissions was 0.634. I felt this was strong enough to do a regression equation with it. However, this correlation coefficient is not strong enough to be completely confident when extrapolating. Therefore, I would suggest that the regression equation is only sufficient for interpolation.
There were some limitations to my data collection. I recorded my data primarily from websites, especially for overseas cars, and for a number of cars which fit all my criteria, they did not necessarily provide all the information I needed and therefore could not be used. As a result of wanting to control as many variables as possible, the scope of cars available for use in my project was also reduced. This therefore reduced the range of available data, and therefore the accuracy of my project. A way to improve this would be perhaps in spending much more time collecting data so as to cover all possible makes and models of cars. Another way this project could be improved or expanded would be comparing the relationships of these three factors between small cars and bigger cars, or perhaps automatic and manual.

16

Bibliography:

Green Vehicle Guide:
http://www.greenvehicleguide.gov.au/GVGPublicUI/QuickCompareWebForm.aspx?CurrentTask=f41ab644-fc06-4695-89ea-22f871793a41

Green Vehicle EPA Guide: www.epa.gov/greenvehicle/

Audi : www.audi.com

Holden: www.holden.com.au/

Peugeot: www.peugeot.com

Subaru: www.suburu.com.au

Volkswagen: www.volkswagen.com.au

17