EFFECTIVENESS OF USING MATHEMATICAL MODELS TO DETERMINE RELATIONSHIPS BETWEEN THE PRICE AND ODOMETER

Table of Contents

[Introduction 2](#_Toc112441967)

[Considerations 2](#_Toc112441968)

[Procedure 2](#_Toc112441969)

[Assumptions and Observations 3](#_Toc112441970)

[Use of Technology 3](#_Toc112441972)

[Results and Discussions 4](#_Toc112441973)

[Evaluate and Verify 6](#_Toc112441974)

[Strengths 7](#_Toc112441975)

[Limitations 7](#_Toc112441976)

[Recommendations 8](#_Toc112441977)

[Conclusion 8](#_Toc112441978)

[Appendices 9](#_Toc112441979)

# Introduction

Teenagers obtaining their license will consider many factors when deciding on their first purchase of a car, majorly depending on the cost. However, the cost will vary depending on safety, performance, and condition. This report will present the findings of a mathematical investigation with the aim of using provided data samples from carsales.com to determine the relationship between the odometer reading and price of a vehicle, specifically a 2015 Mazda 3 Neo.

Due to the size of the data retrieved on carsales.com, many assumptions and observations were made to help remove discrepancies. An appropriate sample of 15 cars were then randomly selected. It was then assumed the odometer reading is the independent variable, and the price is the dependent variable. A scatterplot was then created, showing the price and odometer reading of the selected vehicles. Furthermore, a line of best fit was then plotted to express the correlation between the variables. Due to this, a model was formed by using the gradient formula and the point-slope formula. Finally, the model was tested on a second car, and other electronically developed scatter plots to consider reasonableness, which was then evaluated using strengths and limitations.

# Considerations

- A linear regression model was chosen for analysis of the comparative data sets.

- The independent variable (changed) in this task is the odometer reading (km travelled) as they are the explanatory variables for price. The price is the dependent (measured) variables as they are response variables to the km travelled.

- Due to how specific the nature of the data, a sample size of 15 was chosen to be suitable as it is enough to form a correlation which can be analysed.

|  |  |
| --- | --- |
| It was assumed that …. | It was observed that .... |
| - The odometer displays accurate readings. The identification of the odometer displaying accurate readings will help significantly in finding the correlation between price and the odometer.  - The cars were not modified in any form. Modifications would affect prices of the vehicles, which would degrade reliability of the data collected.  - All the cars in the used car market are chosen to have no previous accident history. Thus, accident history can result in the cost of the car being lower, hindering safety and, unrealistically, could have led to the odometer being tampered with.  - Considering that hundreds of cars are listed on carsales.com, it can be assumed that the location of the car being listed does not have a major effect on its price.  - The vehicles will have no structural damage therefore will not affect the selling price of the vehicle. | - Just because a car is listed on carsales.com for a specific price, doesn’t mean it will sell, and thus could not be worth that listed pricing.  - Used cars have affordable purchase prices. This is because overtime cars depreciate and buying used cars is more economically efficient as much as buying new cars.  - Buying an older model of a car can be like buying a newer model but can come at a significantly reduced price and potentially have a low number of kilometres.  - All the data points collected are a mixture of automatic and manual vehicles, and therefore, will consider the transmission as having no effect on the vehicle.  - Used cars condition worsens over time thus hold less value. |

# Use of Technology

A spreadsheet program was used for a large majority of the investigation to gather results on the data. Microsoft Excel and Desmos were used to create the graphs and tables in this report, and to calculate the following: correlation coefficients, standard deviations, residuals, and sample means.

# Results and Discussions

**Finding the model/equation manually:** The line of best fit was created to express the relationship between the odometer reading (x-axis) and price (y-axis). The line of best fit passes through two estimated data points, allowing for an equal number of data points to be above and below the line. To find the equation/model, two points were estimated A = (19000, 24600)1 and B = (110000,11800) was formed by substituting the two points into the gradient formula ( ). Following the identification of the gradient (-0.1407), Point A and the gradient was inserted into the point slope formula ( to find the equation (). Through observation of the equation, there is a negative gradient (-0.1407) thus there will be a negative relationship between the variables.

**Manual Calculations**

:

-0.1407

***Reasonableness between models:*** ***Figure 1*** shows a manually sketched scatterplot of the Mazda 3 displaying the variable’s odometer reading and price. The sample space is shown in ***Figure 3*.** It is noted that vehicles with higher kilometres tend to be cheaper. Using estimated points, a regression line was added to show the manual equation of Furthermore, excel calculations determine the reasonableness of the manual equation that resulted in the equation . Following the distribution of the data points in ***Figure 1*** it is evident that the data points moderately follow the trendline with a negative correlation. The calculated correlation efficient (𝑟) value of 0.7295 confirms this correlation. Lastly due to the variance in trendlines caused by manual sketching, some points may differ, but it is advisable to propose that both equations are similar, which is further proved in ***Appendix 1***.

***Figure 1:*** *Mazda**Manual Representation of relationship between odometer reading (x-axis) and price (y-axis) shown on a scatter plot:*

Chart, letter

Description automatically generated

***Figure 2:*** *Mazda**Excel Representation of relationship between odometer reading (x-axis) and price (y-axis) shown in the scatter plot:*

***Figure 3:*** *Mazda Sample Space collected from carsales.com*

Table

Description automatically generatedChart, scatter chart

Description automatically generated

**Finding the model/equation using Excel:** The linear regression line is created by which expresses the relationship between the odometer reading and price on Excel. To find the regression line, the standard deviation of x and y were solved to substitute in the gradient formula . Standard deviation shows how dispersed the data is in relation to the mean. Following the identification of the gradient (-0.1409) it was then inserted into to find the equation of

**Excel Calculations**

1. Mean of and values must be found:

2. Standard deviation of “” and “”:

19986.6 y 3296.095

20362

4. Lastly, to find the y intercept, the average of y will be subtracted by b times the average of x (a):

See ***Appendix 1*** for further demonstrations of algebraic techniques

3. Substitute the “” and “” values of standard deviation into the formula to solve for the gradient

Chart, scatter chart

Description automatically generated

***Figure 4:*** *Mazda**Residual Plot*

***Figure 4*** shows a residual plot of the Mazda 3 dataset. A residual plot was produced to validate the assumption of linearity within the data sets. If the distribution of points shows a random scatter across the x-axis, the relationship is most likely linear. ***Figure 4*** shows this, where the points are randomly scattered across the x-axis. Therefore, it can be concluded that a linear model was appropriate. Furthermore, the data points (27000, 42150), (24550,55000), (20250,50800) and (20990, 46000) were identified as outliers due to being positioned an abnormal distance from the rest of the population.

# Evaluate and Verify

***Figure 5:*** *Scatterplot created on Desmos showing the relationship between variables*

Chart, line chart

Description automatically generated

***Considering the reasonableness of models:*** The reasonableness of the Excel model was compared with another electronically developed scatter plot (*Desmos).* Both models have a negative gradient, thus indicating that there is a strong negative relationship between the variables. In the model created on Excel (***Figure 2)***, there is an identical gradient to the model on Desmos (***Figure 5)*** and it has the exact same y-intercept. Even though there are some minor differences between how the models are presented, it can be seen on desmos that both models pass through the same points. Lastly, the distribution of the points above and under the regression line was the same, which led to the same values. Therefore, it can be considered that the model created on Desmos is reasonable.

***Improving the model***: To improve the correlation of the variables, removing outliers can result in a better model. The chosen outliers were removed by determining which points varied the most from the regression line using ***Figure 4.*** Looking at ***Figure 6*** the equation of the line resulted in y= . When comparing this equation of the line to ***Figure 2*,** there was a minor gradient and y intercept difference. The gradient displays the rate of a change of one variable with respect to another, meaning that in this context the correlation got stronger. Further examination showed the increased when compared to the equation with outliers, meaning that the association with the variables improved. However, closer observations determine that there were minor differences in both equations which barely change the correlation. After revisiting the initial observations and assumptions, it should be considered that:

* Location can affect a buyer’s decision.
* Listing prices, may not be worth its listed price
* Condition will vary from car to car.
* Transmission may have an impact on prices.

Therefore, ***Figure 6*** will be disregarded for further evaluation and analysis due to it having insignificant differences between calculations, as well as not accounting for real world limitations.

***Figure 6:*** *Scatterplot created on Excel showing the relationship between variables without outliers*

Chart, scatter chart

Description automatically generated

***Figure 7:*** *Scatterplot created on Desmos testing the relationship between variables on a Toyota Camry*

***Testing the model on a different car:*** The Mazda 3 equation was tested on a similar car (Toyota Camry), which can be deemed similar due to its similar features and similar starting price. When testing the Mazda Equation (**Figure 2)** on the Toyota Camry shown in (***Figure 7*)**, the graph clearly demonstrated a strong negative correlation. Furthermore, both equations show that both values were nearly identical with a difference of 0.005, which furthermore translates the model being suitable. Secondly, both equations have an inversely proportional relationship when the price increases and the km travelled decreases. Lastly, there was a moderate difference between the gradients, which led to the incline of the equations being different. Ultimately the model can be deemed reasonable, but it should be considered that:

* Interpolation of this model will only work on similar bivariate datasets (similar car)
* Errors may have occurred during data collection
* Different car brands may affect price

Graphical user interface, application

Description automatically generatedChart, line chart, scatter chart

Description automatically generated

Text

Description automatically generated with medium confidence

# **Strengths**

- One major strength of the mathematical model is that strong negative correlations were found, which demonstrated inversely proportional relationships, meaning the higher the odometer reading, the lower the price of the car.

- The mathematical model in testing was found to have similar results on other cars. Specifically, the model was tested on a Toyota Camry, which showed similar correlation and showed a similar distribution of points on the trendline.

- Use of technology confirmed accuracy between the mathematical models. Furthermore, Excel improved the reliability of the manual model, which made it reasonable to proceed with the investigation.

- Random sampling (a method of sampling) was used because the selection of sample size is completely random, and it represents accurate sample data and avoids a biased sample that favours one portion of the population.

# **Limitations**

- Taking into consideration an external factor, a variety of locations were used. Certain cars in the dataset could have been affected. For example, cars being listed in urban areas could have a high demand for cars with a lower number of kilometres, causing higher prices than usual. This could explain the reason for outliers.

- The mathematical models used had low sample sizes. Having a decreased sample size can decrease accuracy and increase the margin of error. The data could have been more accurate if a more precise data collection method was used, as well as increasing sample space.

- As the data provided contained outliers, this suggests the model had some inconsistencies in the data collection process, which can impact the correlation between the variables.

- The model will only work on similar bivariate datasets to determine an appropriate correlation between variables. More specifically, a car that has a similar starting price and a similar age will have to be used when testing the model, otherwise the correlation will differ.

# Recommendations

- Considering the sample space, it should be increased. Increasing the sample space will improve accuracy, lower the margin of error in calculations and help justify outliers.

- A small-scale investigation involving the vehicle's previous history would lower the uncertainty established in terms of cost deviance.

- Finding a model that satisfies every car is unrealistic, so it is recommended that a popular car among teenagers be investigated. Consequently, this will improve the overall correlation between the variables.

# Conclusion

The aim of this investigation is to find out how odometer readings affect the price of a vehicle. The analysis was based on data collected on carsales.com and comparisons of mathematical models were made. It can be determined the model showed negative correlations, only suitable for similar cars, and was not an accurate method of determining correlation for all vehicles. Even though the solution is reasonable, due to limitations, it was observed that future recommendations should be made to further refine the model. Specifically, increasing sample space and doing further research on cars will improve the accuracy and reasonableness of the solution.

# Appendices

***Appendix 2:***

***Appendix 1:***