Automobile Price Prediction using Multivariate Linear Regression

Jobb Q. Rodriguez
Ateneo de Naga University
Maryville Subdiv., Brgy. San Felipe
Naga City, Camarines Sur,
+639619116147
jobb.rodriguez22@gmail.com

ABSTRACT

In this paper, the author used Multivariate Linear Regression to identify which experiment performs best in predicting the price given the dataset, *Automobile Dataset* [1]. All sets of features. However, in terms of the optimized cost, the thetas, and the learning rate, Car Performance with Curb Weight vs Price performed best.

CCS Concepts

• Computing methodologies→Machine learning and Modeling and simulation

Keywords

Automobile; Linear Regression; Multivariate; Machine Learning; Supervised

1. INTRODUCTION

The automobile price must be predicted given the set of features. The author's task is to identify which experiment is best in predicting the price given the *Automobile Dataset* [1].

2. RESULTS

2.1 Safety Features vs Price

For the set, Safety Features, the three (3) features used are the "Door Number", the "Wheel Base", and the engineered feature "Car Volume" (car_length * car_width * car_height). In addition, a filler feature containing number ones (1s) for computation purposes. The initial thetas are (0, 0, 0, 0), and the learning rate is 0.001.

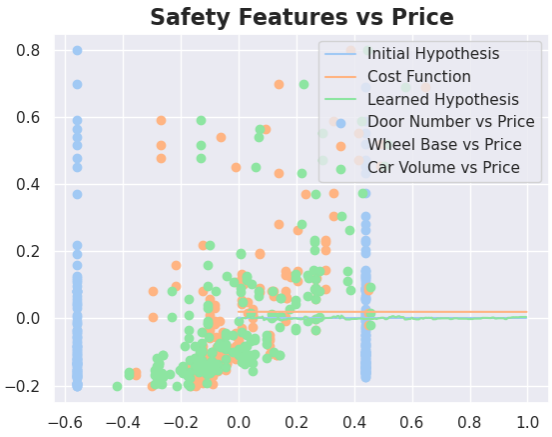

Figure 1. Graph for Safety Features vs Price

The learned hypothesis line passes through some of the scatter plots. Moreover, the former has little improvement when compared to the initial hypothesis line and does not form a pattern towards a specific scatter plot.

Table 1. Final Thetas for Safety Features vs Price

Feature	Theta
Filler	1.2358406980699396e-17
Door Number	0.0008189603336368782
Wheel Base	0.005938011953250316
Car Volume	0.007387462227501046

The cost function line slowly approaches zero (0) and is near zero (0). The optimized cost is 0.019268400489538925.

2.2 Car Size vs Price

For the set, Car Size, the two (2) features used are the "Car Length" and the "Car Width". In addition, a filler feature containing number ones (1s) for computation purposes. The initial thetas are (0, 0, 0), and the learning rate is 0.0011.

Figure 2. Graph for Car Size vs Price

The learned hypothesis line passes through some of the scatter plots. Moreover, the former has little improvement when

compared to the initial hypothesis line and does not form a pattern towards a specific scatter plot.

Table 2. Final Thetas for Car Size vs Price

Feature	Theta
Filler	8.691245555226251e-18
Car Length	0.008104957958698013
Car Width	0.008757741498724138

The cost function line slowly approaches zero (0) and is near zero (0). The optimized cost is 0.019138620255102184.

2.3 Car Build vs Price

For the set, Car Build, the three (3) features used are the "Car Body", the "Car Length", and the "Car Width". In addition, a filler feature containing number ones (1s) for computation purposes. The initial thetas are (0, 0, 0, 0), and the learning rate is 0.0012.

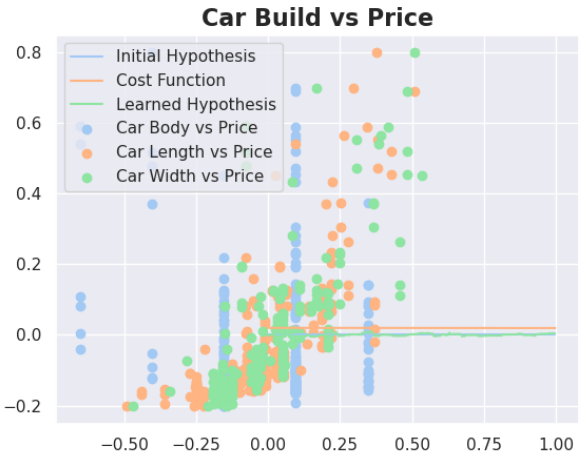

Figure 3. Graph for Car Build vs Price

The learned hypothesis line passes through some of the scatter plots. Moreover, the former has little improvement when compared to the initial hypothesis line and does not form a pattern towards a specific scatter plot.

Table 3. Final Thetas for Car Build vs Price

Feature	Theta
Filler	8.976559333249746e-18
Car Body	-0.0013003194075990262
Car Length	0.008836448752915204
Car Width	0.009546900798837247

The cost function line slowly approaches zero (0) and is near zero (0). The optimized cost is 0.019095329172992188.

2.4 Car Performance vs Price

For the set, Car Performance, the four (4) features used are the "Cylinder Number", the "Engine Size", the "Horsepower", and the "City - MPG". In addition, a filler feature containing number ones (1s) for computation purposes. The initial thetas are (0, 0, 0, 0, 0), and the learning rate is 0.0008.

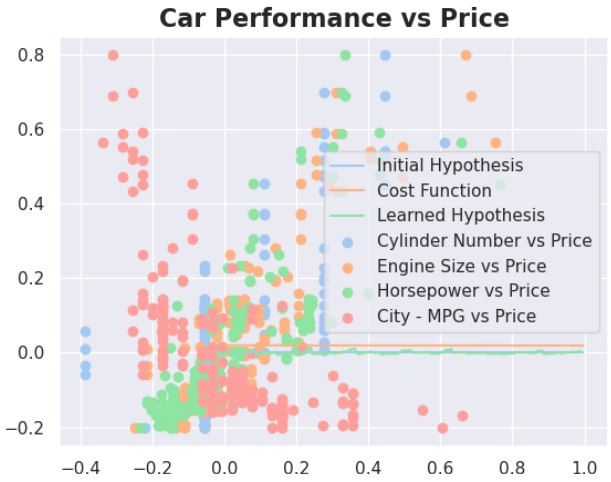

Figure 4. Graph for Car Performance vs Price

The learned hypothesis line passes through some of the scatter plots. Moreover, the former has little improvement when compared to the initial hypothesis line and does not form a pattern towards a specific scatter plot.

Table 4. Final Thetas for Car Performance vs Price

Feature	Theta
Filler	7.48756070695491e-18
Cylinder Number	0.005054895580393018
Engine Size	0.006446921166508406
Horsepower	0.00624362756262352
City - MPG	-0.005839859773595115

The cost function line slowly approaches zero (0) and is near zero (0). The optimized cost is 0.01898592678685277.

2.5 Car Performance with Curb Weight vs Price

For the set, Car Performance with Curb Weight, the five (5) features used are the "Cylinder Number", the "Engine Size", the "Horsepower", the "City - MPG", and the "Curb Weight". In addition, a filler feature containing number ones (1s) for computation purposes. The initial thetas are (0, 0, 0, 0, 0, 0), and the learning rate is 0.0007.

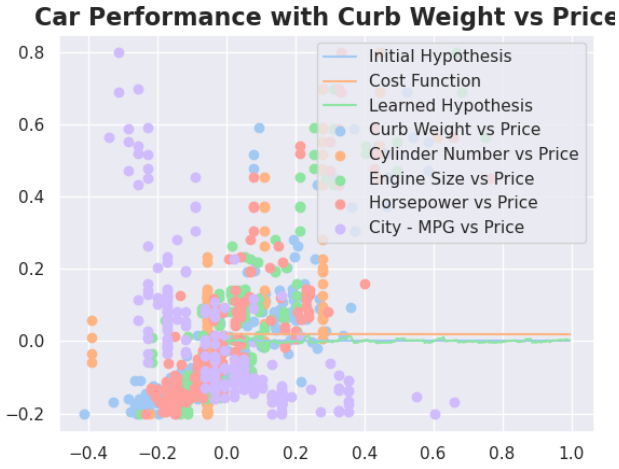

Figure 5. Graph for Car Performance with Curb Weight vs Price

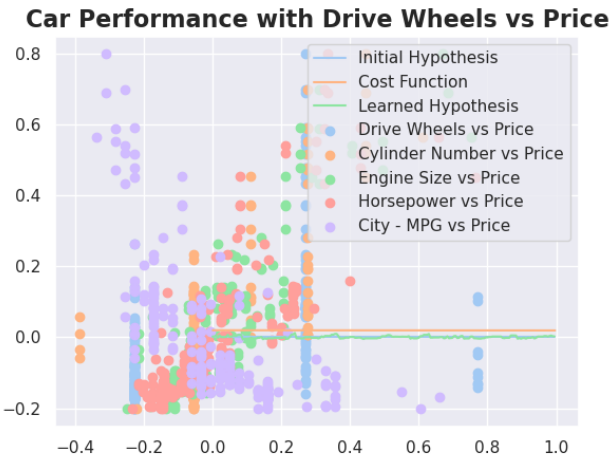

Figure 6. Graph for Car Performance with Drive Wheels vs Price

The learned hypothesis line passes through some of the scatter plots. Moreover, the former has little improvement when compared to the initial hypothesis line and does not form a pattern towards a specific scatter plot.

Table 5. Final Thetas for Car Performance with Curb Weight vs Price

Feature	Theta
Filler	6.469824676003108e-18
Cylinder Number	0.004415042908366098
Engine Size	0.005628120088960888
Horsepower	0.005452121268268306
City - MPG	-0.00509673206379256
Curb Weight	0.0069089959629774845

The cost function line slowly approaches zero (0) and is near zero (0). The optimized cost is 0.01883388008175997.

2.6 Car Performance with Drive Wheels vs Price

For the set, Car Performance with Drive Wheels the five (5) features used are the “Cylinder Number”, the “Engine Size”, the “Horsepower”, the “City - MPG”, and the “Drive Wheels”. In addition, a filler feature containing number ones (1s) for computation purposes. The initial thetas are (0, 0, 0, 0, 0, 0), and the learning rate is 0.0006.

The learned hypothesis line passes through some of the scatter plots. Moreover, the former has little improvement when compared to the initial hypothesis line and does not form a pattern towards a specific scatter plot.

Table 6. Final Thetas for Car Performance with Drive Wheels vs Price

Feature	Theta
Filler	5.629886800653456e-18
Cylinder Number	0.00379507091510805
Engine Size	0.004837935907351297
Horsepower	0.004685016290996561
City - MPG	-0.004379738124539073
Drive Wheels	0.004996186289790507

The cost function line slowly approaches zero (0) and is near zero (0). The optimized cost is 0.0189928674525847.

2.7 Car Size - Order 2 vs Price

For the set, Car Size - Order 2, the four (4) features used are the “Car Length”, “Car Width”, the first engineered feature in Order 2, “Car Length Squared" (car_length^2), and the second engineered feature in Order 2, “Car Width Squared" (car_width^2). In addition, a filler feature containing number ones (1s) for computation purposes. The initial thetas are (0, 0, 0, 0, 0), and the learning rate is 0.0013.

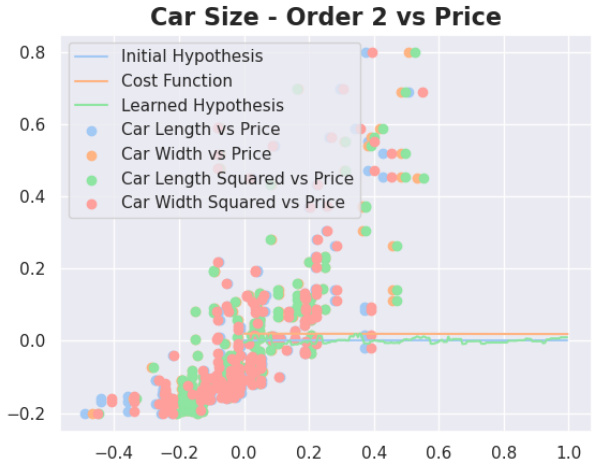

Figure 7. Graph for Car Size - Order 2 vs Price

The learned hypothesis line passes through some of the scatter plots. Moreover, the former has little improvement when compared to the initial hypothesis line and does not form a pattern towards a specific scatter plot.

Table 7. Final Thetas for Car Size - Order 2 vs Price

Feature	Theta
Filler	1.0763950706370362e-17
Car Length	0.00944167202626096
Car Width	0.010215879670176741
Car Length Squared	0.010348433623347631
Car Width Squared	0.009607399542179207

The cost function line slowly approaches zero (0) and is near zero (0). The optimized cost is 0.018562551629153903.

2.8 Car Length - Order 2 vs Price

For the set, Car Size, the two (2) features used are the “Car Length” and the engineered feature in Order 2, “Car Length Squared" (car_length^2). In addition, a filler feature containing number ones (1s) for computation purposes. The initial thetas are (0, 0, 0), and the learning rate is 0.0014.

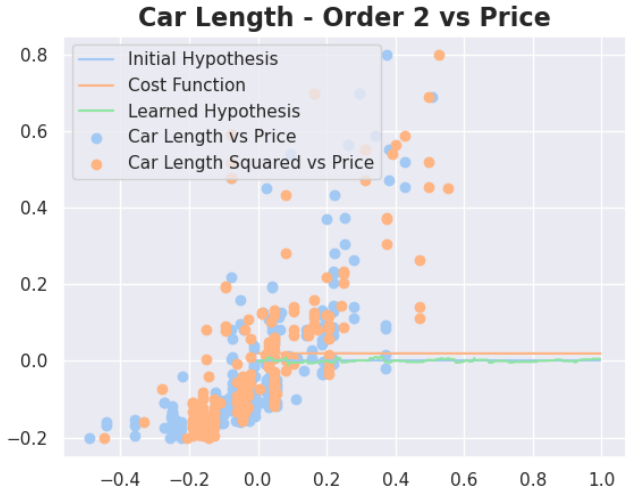

Figure 8. Graph for Car Length - Order 2 vs Price

The learned hypothesis line passes through some of the scatter plots. Moreover, the former has little improvement when compared to the initial hypothesis line and does not form a pattern towards a specific scatter plot.

Table 8. Final Thetas for Car Length - Order 2 vs Price

Feature	Theta
Filler	1.0909755483836058e-17
Car Length	0.01028506488888707
Car Length Squared	0.011259674160278284

The cost function line slowly approaches zero (0) and is near zero (0). The optimized cost is 0.019016372666896434.

3. ANALYSIS

3.1 Thetas, Learning Rate, and Iterations

In terms of the initial thetas, all sets started with zero (0). The sets only differed in the number of zeros in correspondence to the number of features per set.

For Safety Features vs Price, the algorithm prioritized Car Volume and Wheel Base.

For Car Size vs Price, the algorithm prioritized Car Length and Car Width almost equally.

For the Car Build vs Price, the algorithm prioritized Car Length and Car Width.

For the Car Performance vs Price, the algorithm prioritized Engine Size and Horsepower.

For the Car Performance with Curb Weight vs Price, the algorithm prioritized Curb Weight and Engine Size.

For the Car Performance with Drive Wheels vs Price, the algorithm prioritized Drive Wheels and Engine Size.

For the Car Size - Order 2 vs Price, the algorithm prioritized Car Length Squared and Car Width.

For the Car Length - Order 2 vs Price, the algorithm prioritized Car Length and Car Length Squared almost equally.

Table 9. Learning Rates of the Experiments (Arranged from Lowest to Highest)

Experiment	Learning Rate
Car Performance with Drive Wheels vs Price	0.0006
Car Performance with Curb Weight vs Price	0.0007
Car Performance vs Price	0.0008
Safety Features vs Price	0.001
Car Size vs Price	0.0011
Car Build vs Price	0.0012
Car Size - Order 2 vs Price	0.0013
Car Length - Order 2 vs Price	0.0014

Car Performance vs Price has the lowest learning rate, and Car Performance with Curb Weight vs Price has the second lowest learning rate.

All experiments underwent 300 iterations.

3.2 Order and Optimized Cost

Car Size - Order 2 vs Price and Car Length - Order 2 vs Price are the only experiments in Order 2. The rest of the experiments are in order 1.

Table 9. Learning Rates of the Experiments (Arranged from Lowest to Highest)

Experiment	Learning Rate
Car Size - Order 2 vs Price	0.019016372666896434.
Car Performance with Curb Weight vs Price	0.01883388008175997
Car Performance vs Price	0.01898592678685277
Car Performance with Drive Wheels vs Price	0.0189928674525847
Car Length - Order 2 vs Price	0.018562551629153903.
Car Build vs Price	0.019095329172992188
Car Size vs Price	0.019138620255102184
Safety Features vs Price	0.019268400489538925

All optimized costs are acceptableCar Size - Order 2 vs Price has the lowest optimized cost, and Car Performance with Curb Weight vs Price has the second lowest optimized cost.

3.3 Visual Representations

All experiments's learned hypothesis line passes through some of the scatter plots. Moreover, the learned hypothesis line has little improvement compared to the initial hypothesis line and does not form a pattern towards a scatter plot. In addition, all experiment's cost function line slowly approaches zero (0) and is near zero (0).

3.4 Synthesis

In terms of the optimized cost, Car Size - Order 2 vs Price tops the list. However, the feature selection of the thetas is questionable in practical application. Moreover, it has a larger learning rate. Hence, it is not reliable in predicting the automobile's price.

In terms of the optimized cost, Car Performance with Curb Weight vs Price, tops the list after invalidating Car Size - Order 2 vs Price. Moreover, the feature selection of the thetas is applicable in a practical setup. Moreover, it has the second lowest learning rate.

Hence, in comparing the top two (2) experiments in terms of the optimized cost when considering the thetas and the learning rate, Car Performance with Curb Weight vs Price (ranked second) is more practical than Car Size - Order 2 vs Price (ranked first).

4. CONCLUSION

Based on the thetas, the learning rate, and the optimized cost, Car Perform with Curb Weight vs Price performs the best.

5. REFERENCE

[1] Jeffrey Schlimmer. 1965. Automobile Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/automobile