Analysis of Factors that Influence Automobile Price

**Introduction**

For this project, we are interested to see how different characteristics determine the automobile's price, so we will be using linear regression analysis with automobile's price as the response variable. We will use the Automobile Data Set from the Machine Learning Repository at the University of California, Irvine. The original data set consists of 205 observations and 26 variables, however we will be using only 24 variables that relate to the specification of an automobile in terms of various characteristics. For the purpose of our analysis, observations containing missing values for any of the 24 variables were removed, leaving the final dataset of 193 observations of 24 variables:

1. Make (Alfa-Romero, Audi, BMW, Chevrolet, Dodge, Honda, Isuzu, Jaguar, Mazda, Mercedes-Benz, Mercury, Mitsubishi, Nissan, Peugeot, Plymouth, Porsche, Saab, Subaru, Toyota, Volkswagen, Volvo);
2. Fuel type (diesel, gas);
3. Aspiration (std, turbo);
4. Number of doors (four, two);
5. Body Style (convertible, hardtop, hatchback, sedan, wagon);
6. Drive wheels (4WD, FWD, RWD);
7. Engine location (front, rear);
8. Wheel base;
9. Length;
10. Width;
11. Height;
12. Curb weight;
13. Engine type (dohc, l, ohc, ohcf, ohcv);
14. Number of cylinders (eight, five, four, six, three, twelve);
15. Engine size;
16. Fuel system (1bbl, 2bbl, idi, mfi, mpfi, spdi, spfi);
17. Bore;
18. Stroke;
19. Compression ratio;
20. Horsepower;
21. Peak RPM;
22. City MPG;
23. Highway MPG;
24. Price.

**Part 1: Simple Linear Regression Analysis**

First, we will examine how well the price of an automobile can be predicted using only one predictor variable – horsepower. We chose horsepower as the predictor because horsepower measures the power of the car's engine and is an important specification for buyers to consider, as it indicates the car's performance and affects fuel efficiency. Thus, it should explain a lot of variability in price, since a vehicle that has a lot of power is going to be more versatile and should be more expensive.

The scatterplot of the car's price vs horsepower (Fig. 1) shows that, as expected, there is a strong positive relationship between the two variables.

Figure 1. Automobile's Price vs Horsepower



However, the relationship between price and horsepower does not seem to be linear and log-transforming both variables might make a linear model more appropriate since both variables are skewed to the right and logarithmic transformation makes the distributions of the variables more symmetric as shown by the histograms below (Fig. 2). The scatterplot of log-transformed price and horsepower variables (Fig. 3) shows strong positive linear relationship between the two variables, suggesting that linear model is more appropriate for log-transformed variables than it was before transformation.

Figure 2. Histograms of Price and Horsepower before and after log-transformation



Figure 3. Log-transformed Price vs Log-transformed Horsepower



The fitted simple regression line of log of price on log of horsepower is

.

The model suggests that one-percent increase in horsepower is expected to increase the price of the vehicle by about 1.265%. The intercept of the regression line has no meaningful interpretation in this case as the log of horsepower does not take on a value of zero. The residual standard error () of the model is 0.2747 on 191 degrees of freedom, suggesting that the actual log-transformed price of the vehicle can deviate from the true regression line by approximately 0.2747, on average. The R-squared () of the model is 0.7127, which means that log-transformed horsepower explains about 71.27% of variability in the log-transformed price of the vehicle. The below graph (Fig. 4) shows the scatterplot of log-transformed price vs log-transformed horsepower with the fitted regression line superimposed in red and the 90% confidence band for the regression line superimposed in blue.

Figure 4. Log-transformed Price vs Log-transformed Horsepower with regression line (red) and 90% confidence band (blue)



To see whether horsepower has a significant effect on price, we test whether the slope coefficient is different from zero or not at the 5% level of significance ( vs the alternative ). We reject the null hypothesis at 5% level of significance (). The slope coefficient is statistically significantly different from 0, suggesting that horsepower has a statistically significant effect on the price of the vehicle. The 95% confidence interval for the slope coefficient () is (1.151, 1.380), suggesting that we are 95% confident that as horsepower increases by 1%, the increase in the price of the vehicle will be between 1.151% and 1.380%, on average.

The plot of residuals against the predicted values (Fig. 5) shows equally spread residuals around the horizontal line without distinct patterns, which is a good indication that there are no non-linear relationships unexplained by the model and that the homoscedasticity assumption of the linear regression model is satisfied. The normal probability plot of the model residuals (Fig. 6) shows that the distribution of the residuals is slightly skewed as the points form a slight curve instead of a straight line, which suggests that the normality assumption of the linear regression model might be violated.

Figure 5. Residuals vs Fitted Values of the Simple regression Model



Figure 6. Normal Probability Plot of Residuals of the Simple Regression Model



**Part 2: Multiple Linear Regression Analysis**

This part aims at building a linear model to predict the automobile's price using a set of different automobile characteristics.

**Model 1**

First, we fit a linear regression model including three numerical predictor variables: log-transformed horsepower, length and city MPG. Length should be positively related to price as bigger cars tend to be more expensive. City MPG (miles per gallon) should also have an effect on price as fuel efficiency is an important characteristic for most of the buyers.

The fitted regression equation of the first model is

.

The coefficient for log-transformed horsepower suggests that holding length and city MPG constant, one-percent increase in horsepower is expected to increase the price of the vehicle by about 0.942%. The coefficient for length shows that one-unit increase in length is expected to increase the price by 1.8% when horsepower and city MPG stay the same, confirming that bigger cars tend to have higher prices. One mile per gallon increase in city MPG is expected to result in 0.5% higher price, holding horsepower and length of the vehicle constant. The intercept coefficient has no meaningful interpretation in this case as all three variables cannot be zero.

The ANOVA table of the model (Table 1) shows that the total sum of squares (SST) of the model is 50.164 on 192 degrees of freedom; the residual sum of squares (SSE) is 9.355 on 189 degrees of freedom; and the explained sum of squares (SSR) is 40.810 on 3 degrees of freedom.

Thus, the model with log of horsepower, length and city MPG as the independent variables explains about 81.35% of variabiltiy in log of automobile price ().

Table 1. Model 1 ANOVA table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | SS | MS | F | p |
| log(horsepower) | 1 | 35.750 | 35.750 | 722.2882 | < 0.0001 |
| length | 1 | 5.020 | 5.020 | 101.4282 | < 0.0001 |
| city MPG | 1 | 0.039 | 0.039 | 0.7961 | 0.3734 |
| Residuals | 189 | 9.355 | 0.049 | - | - |

To see whether this model is significantly better than a model with no independent variables, we perform an F test of overall significance, testing vs the alternative . We reject the null hypothesis at 5% level of significance (). The model fits the data statistically significantly better than an intercept-only model, which means that horsepower, car length and city MPG have a significant effect on price.

Finally, we test whether city MPG is helpful in prediciting the automobille's price, given that log of horsepower and car length are in the model ( vs ). We cannot reject the null hypothesis at 5% level of significance (). The coefficient for city MPG is not statistically significantly different from 0, suggesting that city MPG is not helpful in explaining the price of the vehicle once we account for horsepower and the length of the car.

**Model 2**

For the second multiple regression model, we include log-transformed horsepower, fuel type (diesel or gas) and their interaction term as predictor variables for log-transformed price.

We incorporate the categorical variable fuel type into the model by defining an indicator variable *gas*, which equals 1 if the car's fuel type is gas, and 0 if the car's fuel type is diesel.

The fitted regression equation of the second model is

The coefficient for gas variable suggests that cars with gas fuel type are expected to be 14.9% more expensive compared to diesel cars when the horsepower of the car is 1 (or log of horsepower is 0). However, this interpretation has no practical meaning as a car cannot have horsepower of 1. The other coefficients suggest that for diesel cars, 1% increase in horsepower is expected to increase the price of the vehicle by about 1.478% while for gas cars, 1% increase in horsepower is expected to increase the price by about 1.329% (1.478 – 0.149 = 1.329).

However, the coefficient for the interaction term between log-transformed horsepower and fuel type is not statistically significant as shown by testing the hypotheses vs . We cannot reject the null hypothesis at 5% level of significance (. This suggests that the effect of horsepower on price is not significantly different for gas and diesel cars and the interaction term can be dropped from the model.

**Model 3**

For the third and final model, we build the "best" regression model using all available variables in the data set, indicating various car characteristics. We use AIC as the criterion to obtain the "best" model using "backward elimination" procedure. 18 out of 23 predictors are included in the final model (AIC = -828.045): Make, Aspiration, Number of doors, Body style, Drive wheels, Wheel base, Length, Width, Height, Curb weight, Engine type, Number of cylinders, Engine size, Fuel system, Bore, Compression ratio, City MPG, Highway MPG.

These variables together explain about 96.96% of variability in log-transformed automobile price ().

**Conclusions**

The aim of this project was to examine car characteristics affecting the automobile price. The results of linear regression analysis showed that log-transformed horsepower alone can explain as much as 71.27% of variability in the log-transformed price of the vehicle and that the effect of horsepower on price is the same for gas and diesel cars. The final model showed that various different car characteristics contribute in determining the price of the vehicle and thus it is difficult to determine the most important factor affecting vehicle price.

One shortcoming of this study is that the data used is from 1985 Ward's Automotive Yearbook. Using data for more recent years might lead to different results and different car characteristics might be important in determining the price as automobiles have changed quite a lot since 1985.