Analysis of Factors Influencing a Car's Price

Abstract

The following code is written to analyze the automobile data taken obtained by UCI [Link: https://archive.ics.uci.edu/ml/datasets/Automobile]. The main focus of the analysis is on how price is affected by certain factors like fueltype, mpg on highways, weight of a car, among others. Analysis techniques include simple, multiple and logistics regression, correlation analysis, CI analysis and bootstrapping, as well as descriptive statistics.

Regression Results

Weight showed highest absolute correlation with price, thus I investigated the relationship further and came to the conclusion that each additional unit of weight is, according to our simple linear model, associated with a 10.9 unit increase in price. This model appeals to our intuition that larger vehicles are likely more expensive to produce and, in turn, price more highly for the consumer. Of course, a simple linear model doesn't prove such a causal relationship, but this linear association appeals to reason.

Due to problematic behaviour in the diagnostic plots, I further conducted a multiple regression on log-variables. Overall, I can say, based on the adjusted R squared, the diagnostic plots as well as the reduced error that our regression model improved. However, I are still facing heteroscedasticity and dependence issues, thus our results need to be taken with a grain of salt. But a further development of our model is beyond the scope of the project.

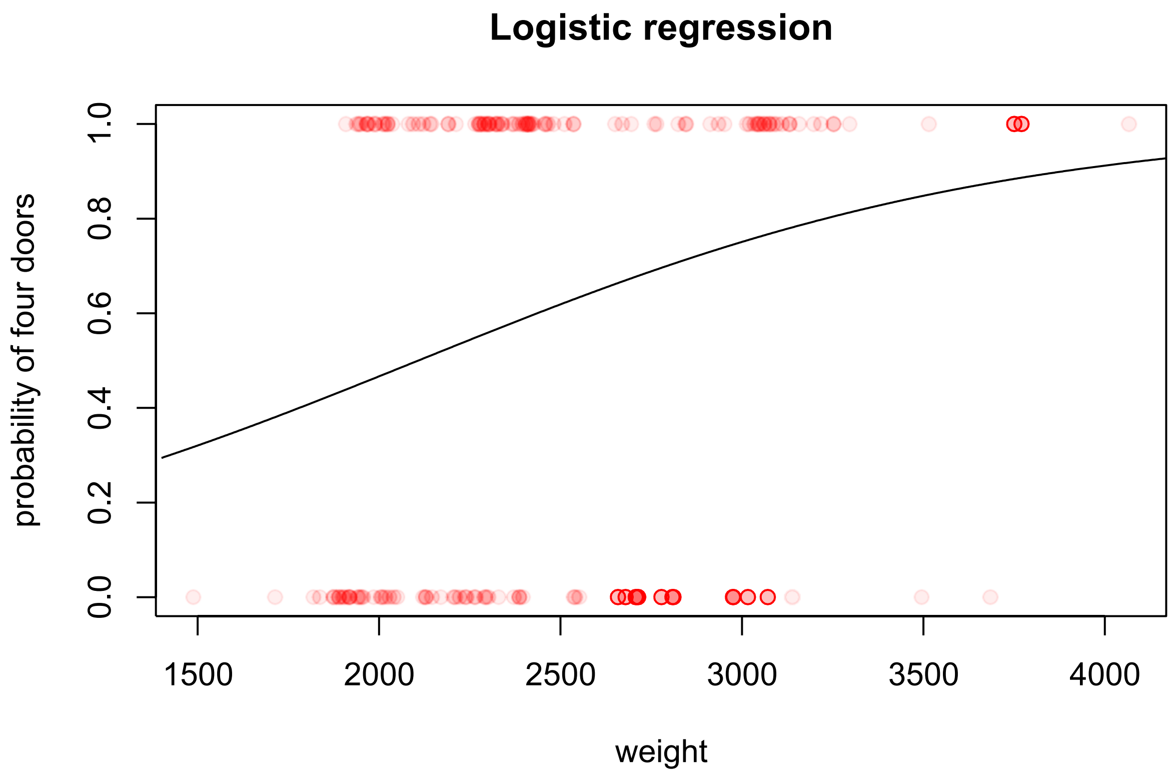
Confidence Interval for Price depending on Fuel Type

I examined how the CI of price was dependent on the fuel type of a car. Interestingly, the t boostrap yields a confidence interval that significantly differs from the one I found via the percent bootstrap method. In both cases, however, the result indicates that the mean of diesel cars is, with high certainty, greater than the mean of gas cars. Chihara and Hesterberg indicate that the t bootstrap method is more accurate. Again, note that the t bootstrap CI is asymmetric, while the t bootstrap CI is not. This serves reminds us of the importance of analyzing the distribution of our data when deciding on confidence interval methods.

Logistic Regression

Lastly, I performed a logistic regression. One would suspect that manufacturers are more apt to equip larger cars with more doors than they are smaller cars. Thus, I think it makes sense to model the probability of a car having four doors (sybolized by 1) instead of two (symbolized by 0), by the weight of the car. Therefore, I conducted a logistic regression on door number against car weight.

```
##
## Call:
## glm(formula = doors.vect ~ weight, family = binomial)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
1
## -2.0412 -1.1448 0.7269 1.0438 1.2832
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.6051489 0.9497715 -2.743 0.00609 **
## weight 0.0012362 0.0003912 3.160 0.00158 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 214.34 on 158 degrees of freedom
## Residual deviance: 202.94 on 157 degrees of freedom
## AIC: 206.94
##
## Number of Fisher Scoring iterations: 4
```



Both visually and computationally, the suspicion is confirmed.