bst260_final

Ellen Chen

12/16/2022

Introduction

The dataset used in this project was from Kaggle. The data includes the price and features of vehicles, including company, car types, carlength, carwidth, symboling, citympg, highwaymph, etc. According to the United States of Environmenal Protection Agency (EPA), for gasoline vehicles, the label shows city, highway, and combined MPG values. The conbined MPG, which will refer as "Fuel Economy" in this project, is calculated by weighting the city value by 55% and the highway value by 45%.

$$\text{Fuel Economy} = 0.55 \times \text{citympg} + 0.45 \times \text{highwaympg}$$

Since the EPA did not state how to calculate fuel economy for diesel vehicles, I will filter diesel vehicles out of the data for the analysis. The category of symboling was used to indicate the degree of safety. Cars are initially assigned a risk factor symbol associated with its price. The symbol is then adjusted by the safe/risky conditions. A value of $+3$ suggusts the auto is risky while -3 suggests the auto is safe.

By navigating the data, the distribution of price is right-skewed. Most of the prices that had been reported were lower than 15k (Figure 1). The top 3 companies that had been reported were Toyota, Nissan, and Mazda (Figure 2). The most two commonly reported car types were hatchback and sedan (Figure 3). The four companies having vehicles with prices higher than 30k were BMW, Buick, Jaguar, and Porsche. Considering car types, all the wagons were reported with prices lower than 20k, but the other 4 types did not show a clear pattern with regard to price (Figure 4).

Results

Primary Analysis

To analyze if fuel economy is associated with the price of vehicles, a simple linear regression model was firstly created, considered the fuel economy as the only predictor.

The fitted model is

$$\hat{Y}_{price} = 37.98 - 0.92X_{fuel.economy}$$

The adjusted R-squared for the simple linear regression model is 0.536.

For each one-unit increase in fuel economy, it is expected to have a -0.92 decrease in price(k) on average, with $p < 0.001$ (Table 1). The fuel economy was shown to be negatively correlated with the price. However, the plot suggests a quadratic relationship between price and fuel economy (Figure 5).

I then introduced a quadratic term into the model, the fitted model was modified to be

$$\hat{Y}_{price} = 80.05 - 3.96X_{fuel.economy} - 0.052X^2_{fuel.economy}$$

The association between fuel economy and price is nonlinear, because the quadratic term has a non-zero slope with $p < 0.001$ (Table 2). The quadratic curve fits better compare to the linear line (Figure 6 and Figure 7). However, neither car type nor symboling showed a clear pattern considering the relationship between fuel economy and price.

In addition, the adjusted R-squared for the quadratic model is 0.701 which is greater the the value of the adjusted R-squared for the simple model. This also indicates that the introduction of the quadratic term improves the model mroe than would be expected by chance.

Secondary Analysis

The secondary analysis aims to study the the association between car types and the safety of vehicles. Association tests, including chi-square analysis, odds ratio, and confidence interval, were utilized to analyze if the car type is associated with safety. Considering the sample size and the distribution of the number of car types in the data, I used a subset of the original data that only contains two car types, hatchback and sedan, which were the best sellers among the five car types.

Based on the result of the chi-square analysis, car types is associated with the safety, with $p < 0.001$. The bar plot also showed a difference in the proportion of safe/risky cars between hatchback and sedan, which was consistent with the result of chi-square analysis (Figure 8).

Next, the odds ratio was calculated to quantify the association between car types and safety. The probability of being safe was divided by the probability of being risky for both hatchback and sedan. The odds of sedan was then divided by the odds of hatchback for the calculation of odds ratio. The result demonstrated that sedan is 9.70 more likely to be safe compared to hatchbacks, with 95% CI [4.23, 22.22].

Conclusion

The fuel economy is found to be associated with the price of vehicles by regression analysis. By adding the quadratic term, the regression model fits better to the original data. However, further analysis needs to carefully measure if the quadratic model performs well compared to the simple linear regression model and how much it improves from the simple model. Confounders and effect modifiers should also be considered and measured in the model.

Taking advantage of association tests, car type is found to associated with the degree of safety of vehicles. However, the analysis only included hatchback and sedan. Further research could include more car types and probably other ways to measure the safety of vehicles besides symboling.

References

EPA: Text Version of the Gasoline Label

Kaggle Database

Appendix

Figure 1: The distribution of price

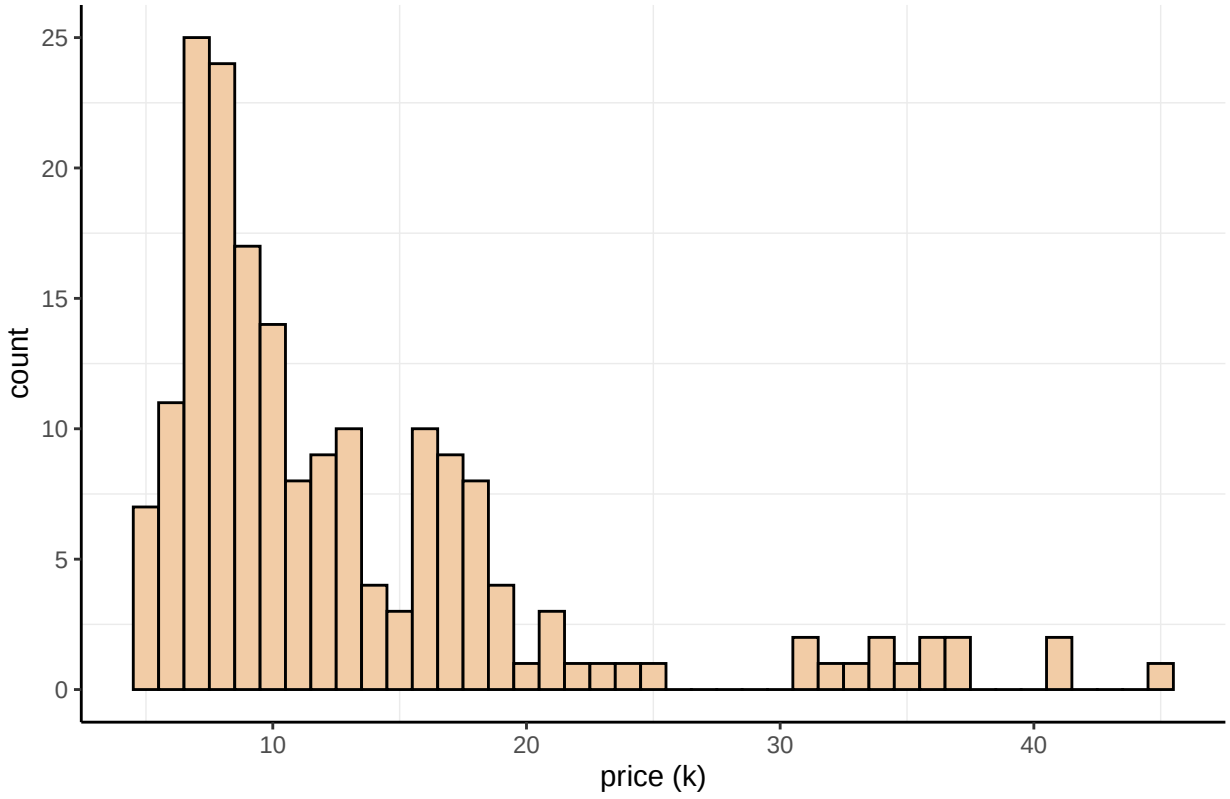


Figure 2: The distribution of company

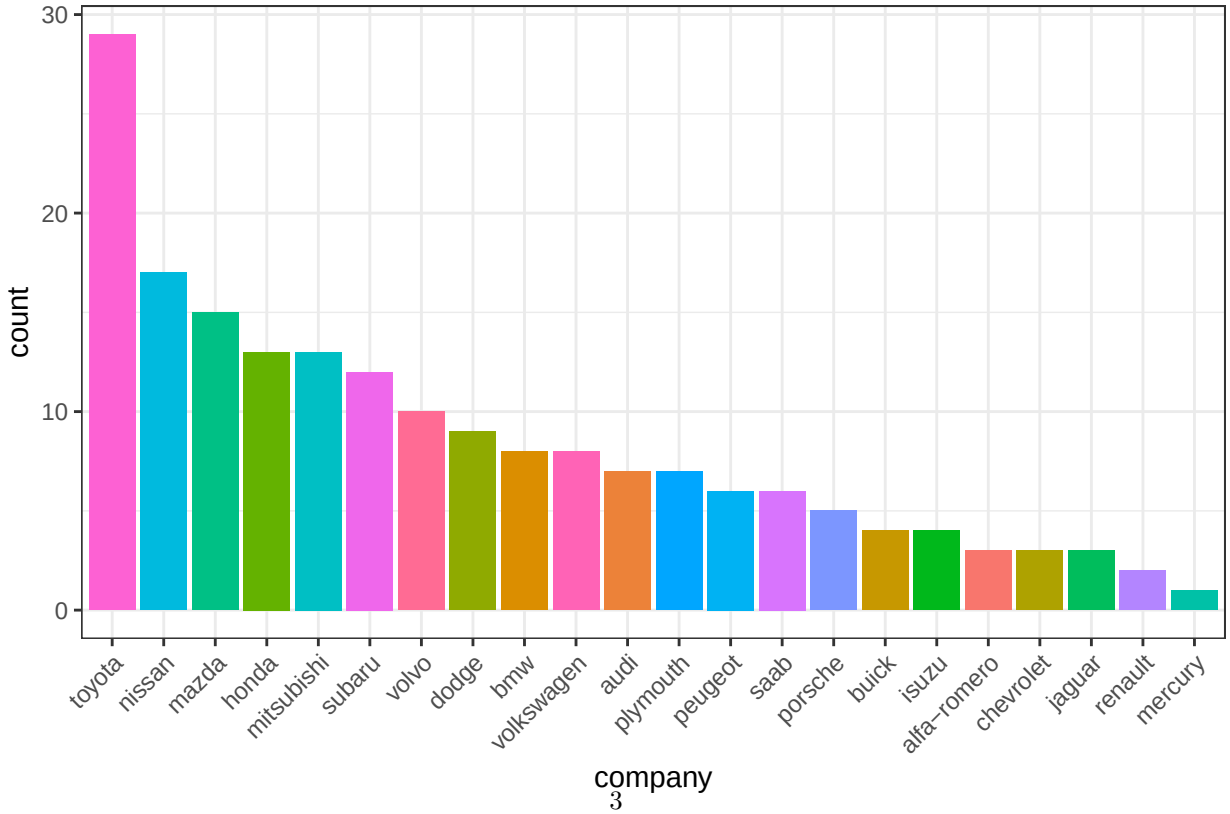


Figure 3: The distribution of car types

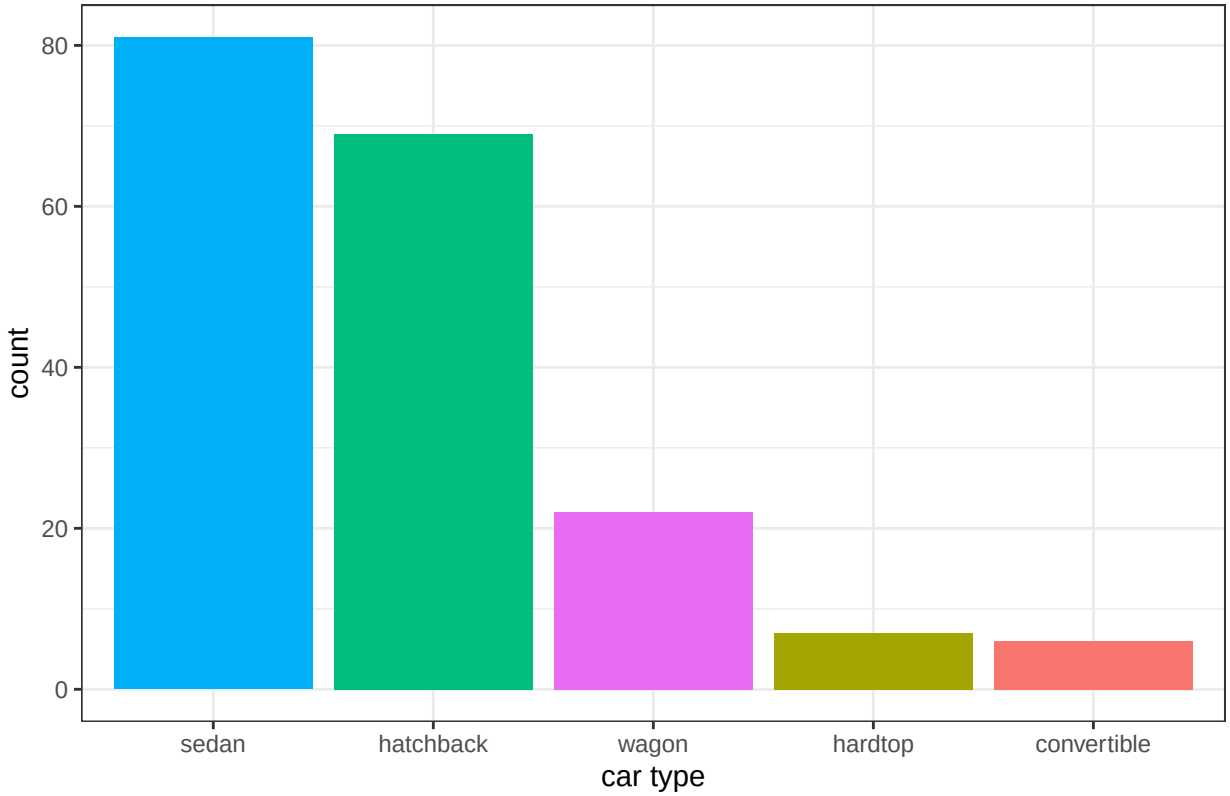
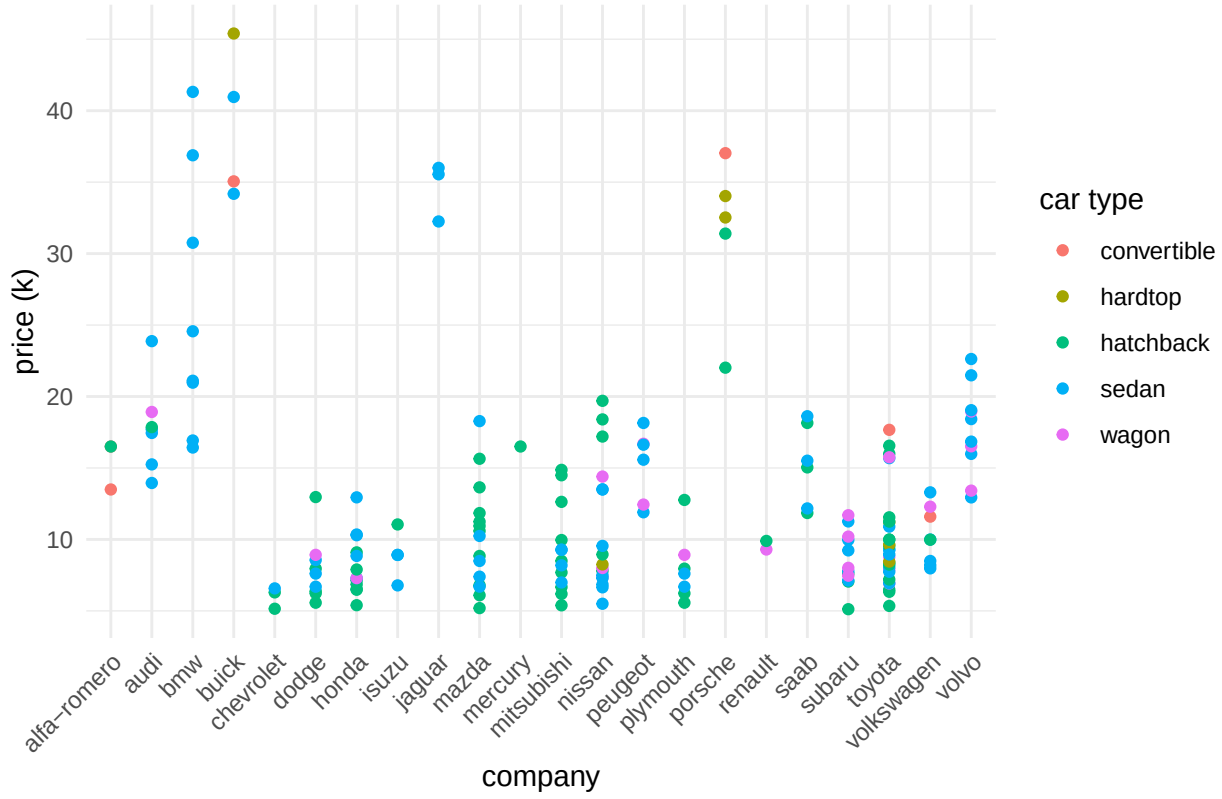


Figure 4: Price vs. Company

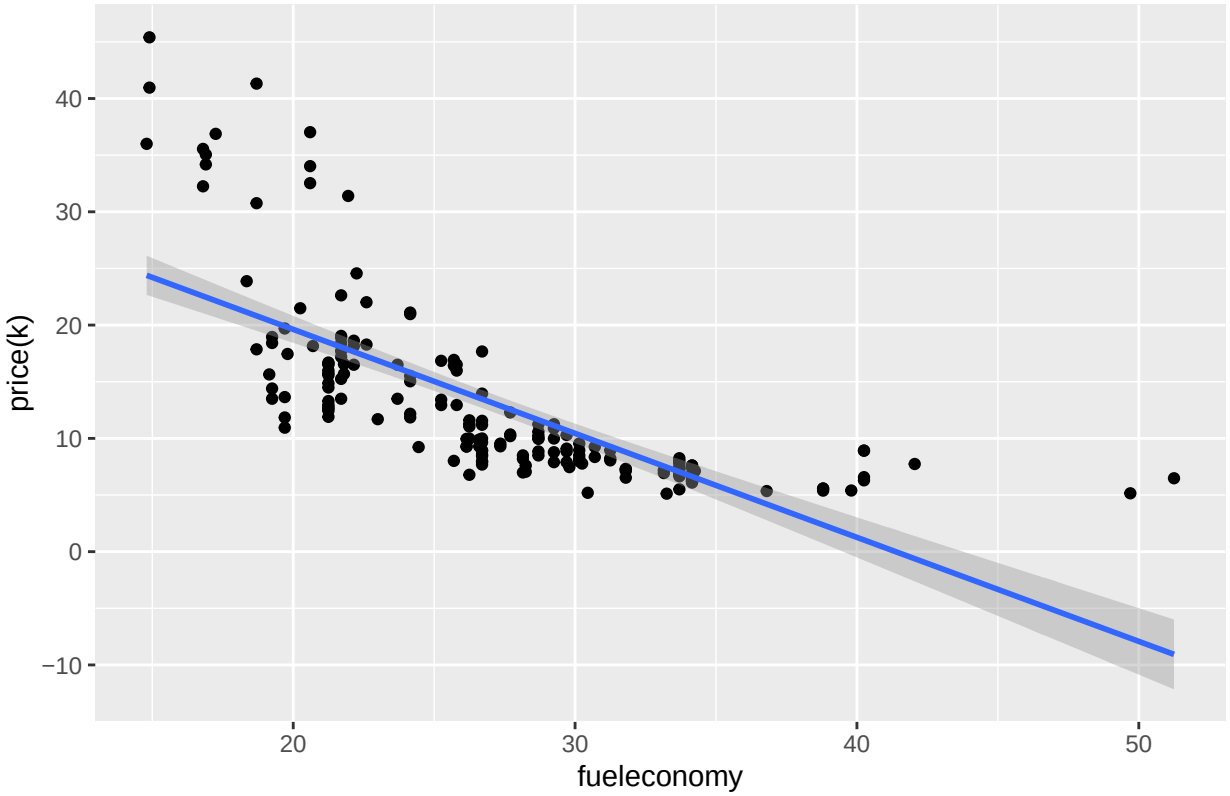



```
## [1] "Table 1 refers to the following output of the simple linear regression model."

##
## Call:
## lm(formula = price ~ fueleconomy, data = car)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9511 -3.1389 -1.1470  0.8599 21.0974
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.98112    1.75573   21.63   <2e-16 ***
## fueleconomy -0.91802    0.06282  -14.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.439 on 183 degrees of freedom
## Multiple R-squared:  0.5385, Adjusted R-squared:  0.536
## F-statistic: 213.5 on 1 and 183 DF,  p-value: < 2.2e-16

## `geom_smooth()` using formula 'y ~ x'
```

Figure 5: The simple linear regression model (price ~ fueleconomy)




```
## [1] "Table 2 refers to the following output of the quadratic model."

##
## Call:
## lm(formula = price ~ fueleconomy + I(fueleconomy^2), data = car)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3162  -2.5544   0.2935   1.4985  17.0957
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      80.047824   4.404181   18.18   <2e-16 ***
## fueleconomy      -3.960749   0.305976  -12.95   <2e-16 ***
## I(fueleconomy^2)  0.052153   0.005173   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.369 on 182 degrees of freedom
## Multiple R-squared:  0.7039, Adjusted R-squared:  0.7006
## F-statistic: 216.3 on 2 and 182 DF,  p-value: < 2.2e-16
```

Figure 6: The quadratic model, color = car type

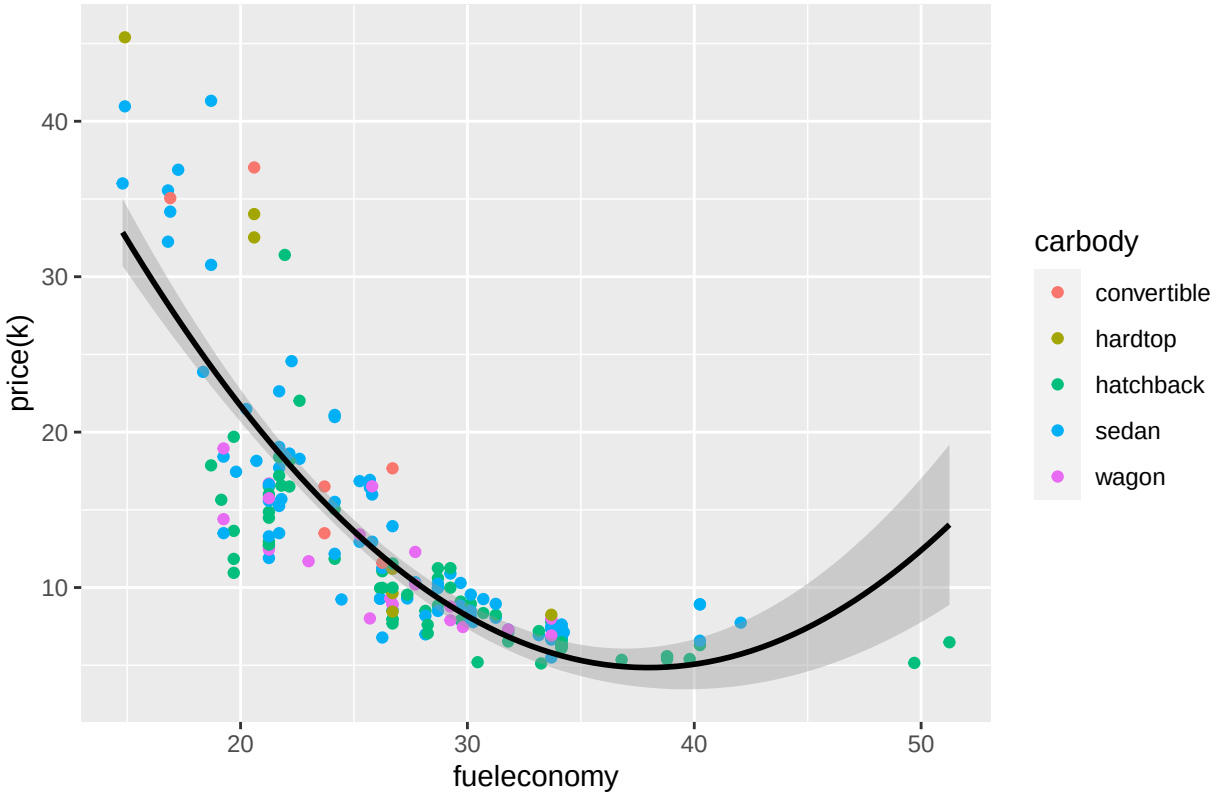
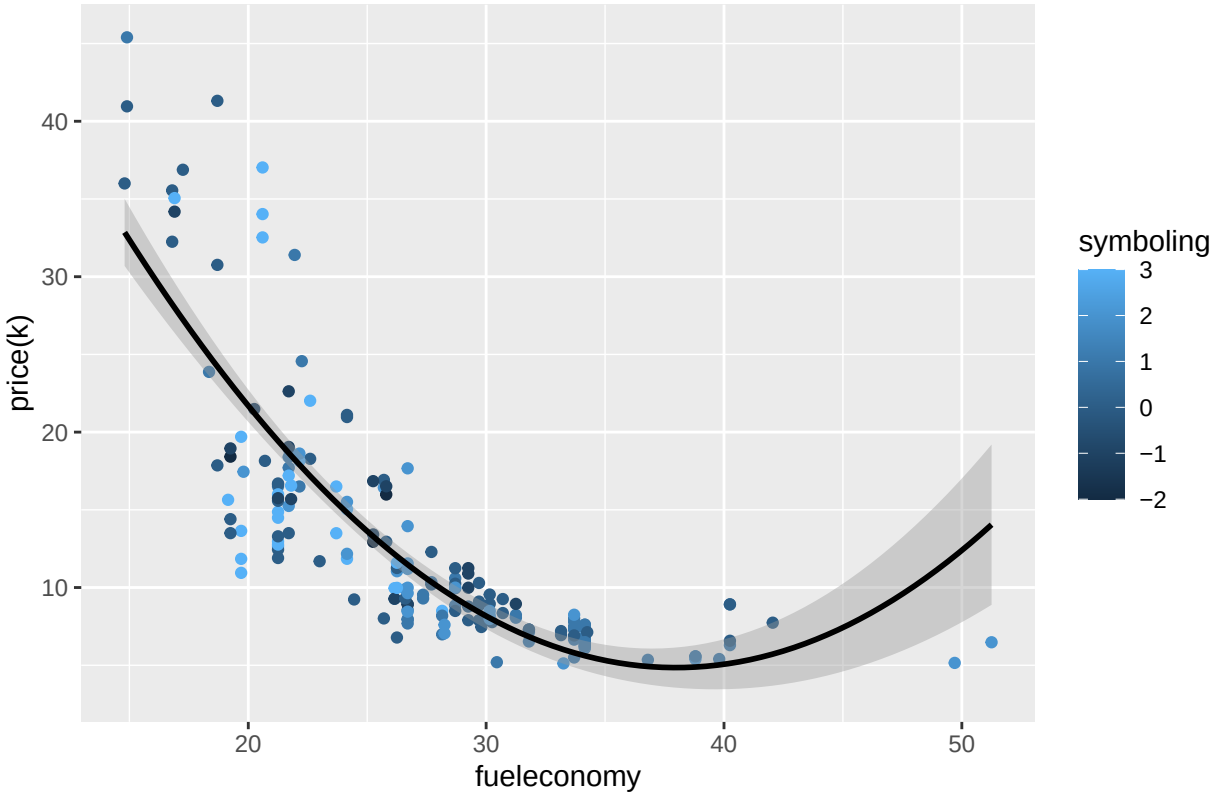
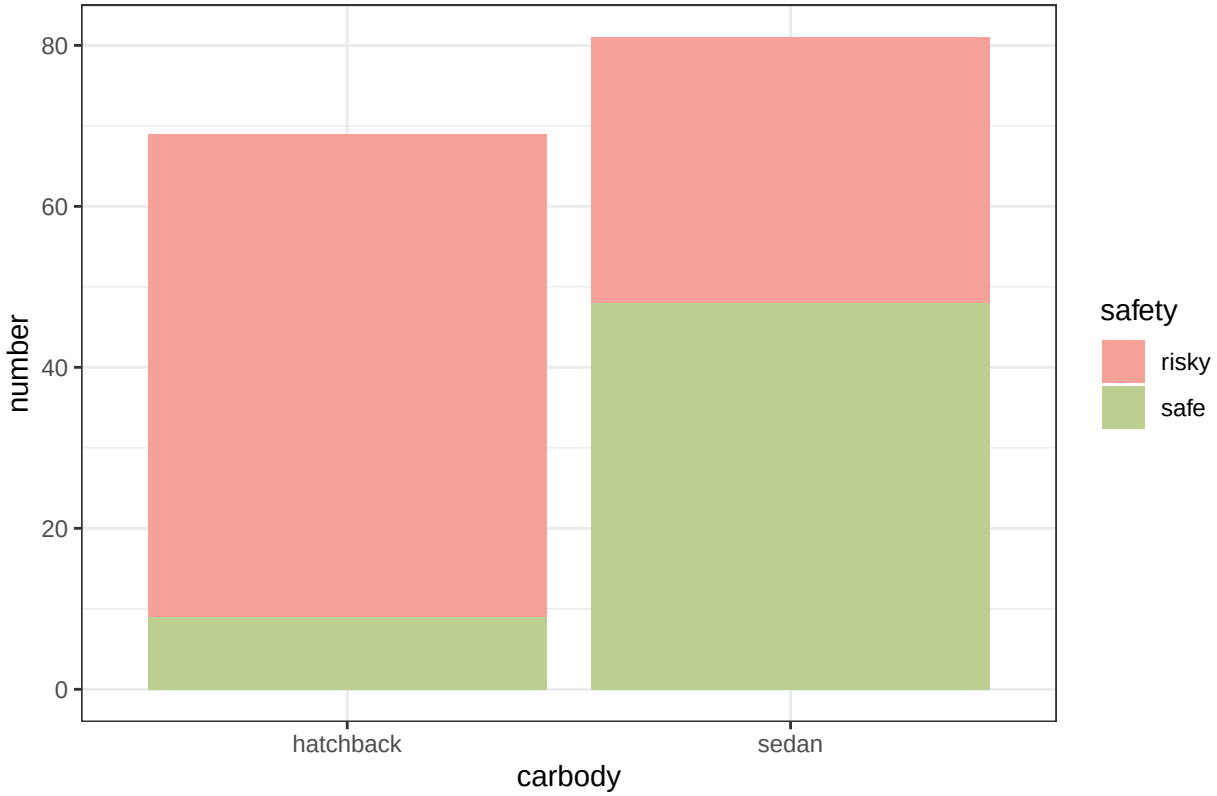


Figure 7: The quadratic model, color = symboling




```
## [1] "The p-value fo the Chi-square analysis is 1.66896043834047e-08"
```

Figure 8




```
## [1] "The odds of hatchback is 0.15"

## [1] "The odds of sedam is 1.45454545454545"

## [1] "The odds ratio is 9.6969696969697"

## [1] "The lower bond of 95% confidence interval is 4.2324587674852"

## [1] "The upper bond of 95% confidence interval is 22.2166892744047"
```