

Toyota Case Study

Trevor Isaacson, Runze Jiang, Adam Kiehl

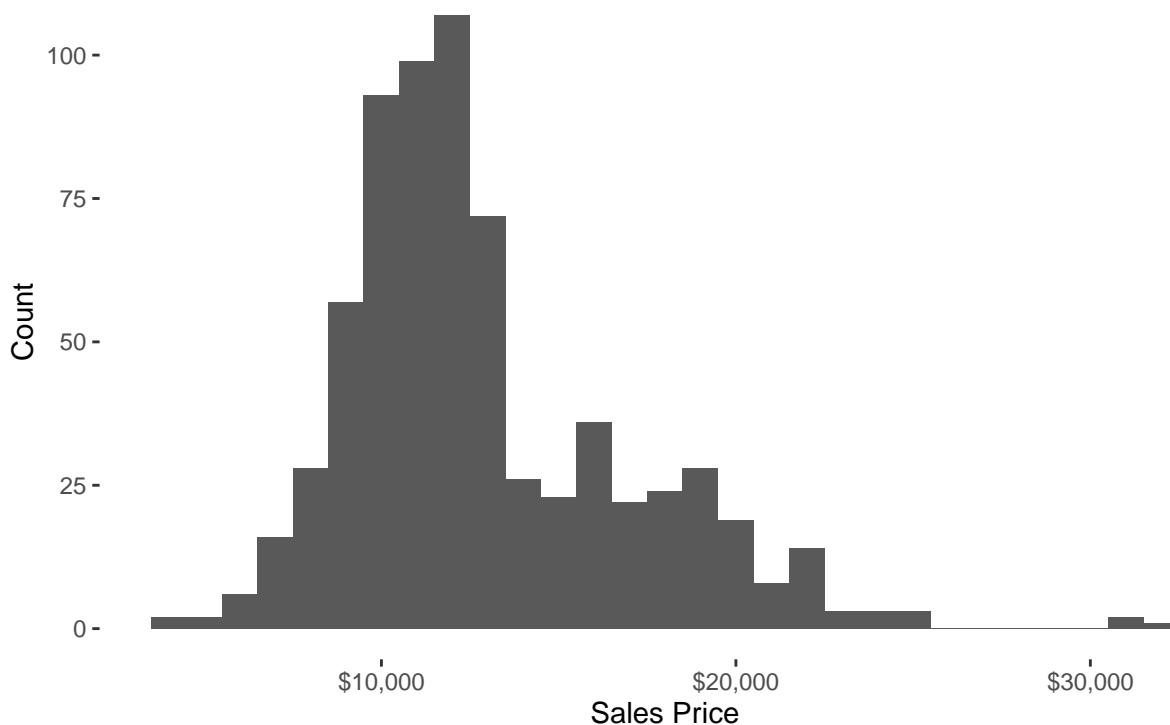
2022-10-12

Introduction

Consumers shopping for new cars often have the option to trade in their current car for money towards their new one. Dealerships, in turn, sell the used cars at a slightly higher price. To do this effectively, the dealership must be able to accurately estimate the expected price that it can sell each used car for based only on characteristics of the car.

Distribution of Used Toyota Corolla Sales Prices

Figure 1



A Toyota dealership was interested in fitting a regression model to estimate `Price` (Figure 1) using 16 different predictor variables and a data set of 694 used Toyota Corolla sales. Numeric predictors included `Age`, `KM` (odometer reading), `HP` (horsepower), `CC` (cylinder volume), `QuartTax` (quarterly road tax), `Weight` and `Period` (guarantee period). Categorical variables included `Mfg_Month` (manufacturing month), `Mfg_Year` (manufacturing year from 1999-2004), `Fuel_type` (CNG, Diesel, or Petrol), `Metallic`, `Automatic`, `Doors` (3, 4, or 5), `Gears` (5 or 6), `Guarantee` (manufacturer's guarantee), and `BOVAG` (BOVAG guarantee). The

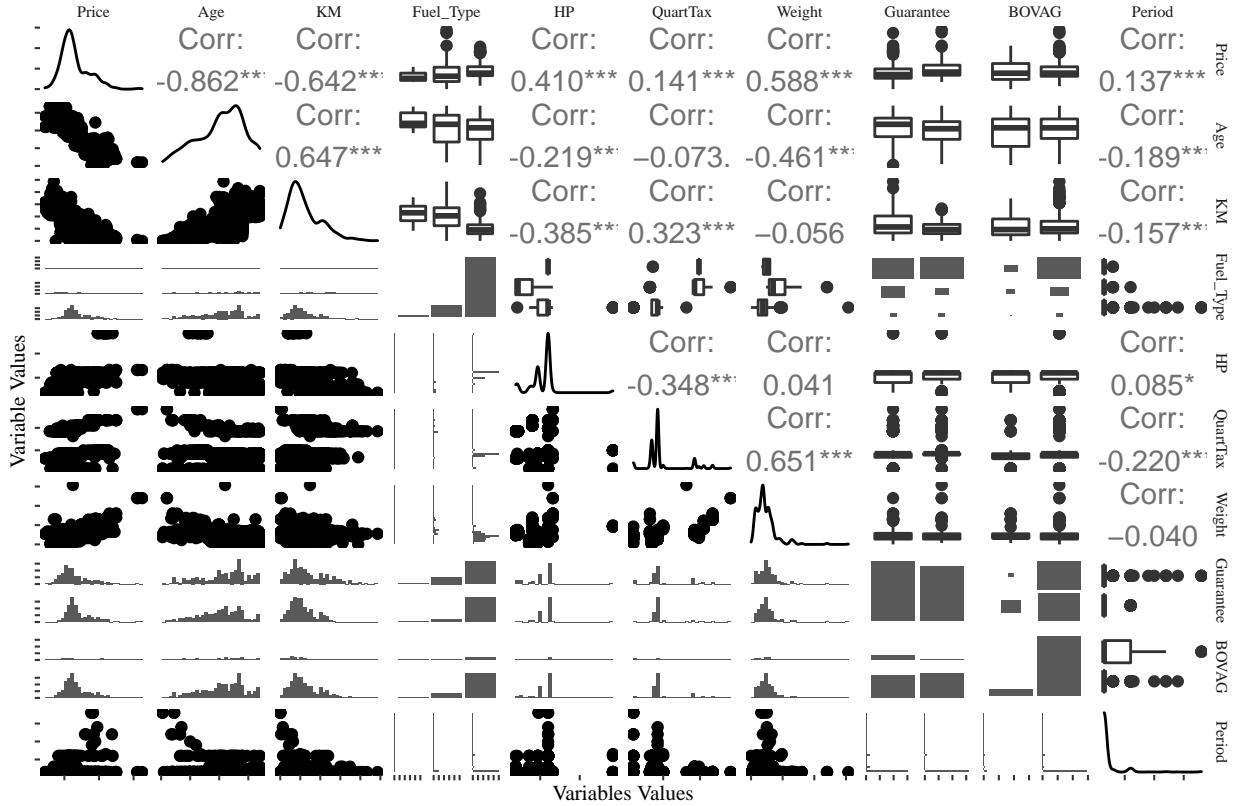
Table 1: Summary Statistics

| Variable | Mean | Std. Deviation | Minimum | Maximum |
|-----------|----------|----------------|---------|---------|
| Age | 41.31 | 15.77 | 1 | 68 |
| BOVAG | 1.90 | 0.30 | 1 | 2 |
| Fuel_Type | 2.80 | 0.44 | 1 | 3 |
| Guarantee | 1.48 | 0.50 | 1 | 2 |
| HP | 102.49 | 16.95 | 69 | 192 |
| KM | 60614.31 | 41309.23 | 1 | 243000 |
| Period | 4.03 | 3.81 | 3 | 36 |
| Price | 12928.45 | 4012.51 | 4350 | 32500 |
| QuartTax | 96.47 | 48.92 | 19 | 283 |
| Weight | 1090.51 | 62.13 | 1000 | 1615 |

Cylinders predictor was removed from the original data set because all recorded cars had 4 cylinders, and all categorical predictors were made into factors in R. No other cleaning was performed on the data.

A set of practical predictors of interest were chosen based on contextual knowledge and exploratory data visualizations to use for model fitting separately of any formal variable selection techniques. Intuitively, older cars (indicated by **Age** and **KM**) should be worth less and larger and more powerful cars (indicated by **Weight** and **HP**, respectively) should be worth more. It was assumed that cars that take different fuel types (**Fuel_Type**) are likely different kinds of cars and would likely vary in price in some way. Finally, **Guarantee**, **BOVAG**, and **Period** are financial-based incentives that should also intuitively lead to a higher car price. Summary statistics (Table 1) and plots (Figure 2) for these predictors of interest are shown below.

Paired Plots and Correlations (Figure 2)



Analysis

The relationships between many of the predictors and the response variable of interest, car price, was largely unknown. Aside from a selection of predictors deemed to likely be important based on contextual knowledge, it was not known which predictors should be included in a regression model. Therefore, an array of models were fit using various transformations, variable selection methods, and prior assumptions. All models were fit with the `stan_glm` function in the `rstanarm` package using a Bayesian framework that incorporates prior belief into traditional regression by use of prior distributions. After fitting, all models were assessed for explanatory power, overfitting, and violations of model assumptions by comparing R^2 and R^2_{LOO} and by generating QQ-Normal plots for residuals and residual vs fitted values plots.

To begin, a full model (1) was fit using the default `stan_glm` priors ($N(0, 6.25)$) to serve as a comparative baseline for future models. The regression intercept was estimated to be $\beta_0 = -4,173.88$ and can be interpreted as the expected price (in dollars) of 3-door, manual, non-metallic, CNG-fueled car that was manufactured in January of 1999 with no manufacturer or BOVAG guarantee and a value of 0 for all numeric predictors included in the model. The regression coefficients, β_j , could be interpreted as the expected difference in expected car price associated with a one unit difference in the predictor (or associated with whether or not an observation assumes that specific level of a categorical predictor), with all other predictors remaining constant. These interpretations are unreasonable in a practical context and it makes no sense that the price of a car could be estimated to be a negative value. Therefore, another full model (2) was fit, also using the default `stan_glm` priors, but now with scaled numeric predictors and a log-transformation of car price. This was done to make the interpretation of coefficients more practical and comparable, and to restrict predictions of price to positive values. There may also be potential for modeling benefits from scaling. Then, regression coefficients for numeric predictors could be interpreted as the expected percent difference in expected car price associated with a one standard deviation increase in the value of the predictors, with all other predictors remaining constant.

The full models fitted utilized 32 predictors (16, not including dummy variables) for modeling. Using a large number of predictors can be damaging to a model due to redundancies in predictor information (multicollinearity) and the potential for overfitting. Several different variable selection methods were applied to the car prices data set to determine which predictors were most closely related with car price. First, a Bayesian variable selection method called a horseshoe prior was used to systematically exclude unimportant predictors from the model. A horseshoe prior places a large amount of prior mass near 0, forcing predictors to be highly related with the response to be included in the model. This model (3) determined 14 predictors (12, not including dummy variables) to be significant in predicting car price: `Age`, `Mfg_Month`, `Mfg_Year`, `KM`, `Fuel_Type`, `HP`, `Metallic`, `QuartTax`, `Weight`, `Guarantee`, `BOVAG`, and `Period`. A follow-up model (4) was fit using only these selected predictors, with only an inconsequential loss in R^2 . Next, a model (5) was fit using a variable selection method called LASSO that penalizes a model for producing large coefficient estimates, forcing it to be selective with how it chooses estimates (equivalent to using a Laplace prior). This model determined 15 predictors (12, not including dummy variables) to be significant in predicting car price. These were the same predictors chosen by the horseshoe prior model previously fit. A follow-up model (6) was fit using only these selected predictors, again with only an inconsequential loss in R^2 (DO WE NEED THIS MODEL?? ISNT IT THE SAME AS 4??).

Thirdly, a more contextual approach was taken to variable selection and the predictors selected previously based on practical knowledge and data visualizations were used for modeling. These predictors were selected independently of the results of the horseshoe and LASSO selections, but aligned closely with these selections nonetheless. First, a model (7) was fit using these predictors and the `stan_glm` default priors. A follow-up model (8) was then fit using weakly informative priors (Table 2) informed by prior knowledge and exploratory data visualizations. This model showed slightly less evidence of overfitting than the previous model but was otherwise similar. Finally, two models (9) & (10) were fit to explore potential interactions of interest. The first of these models included an interaction between `Age` and `KM`, and the second included an interaction between `Weight` and `HP`.

Table 2: Prior Parameters

| Variable | Prior Mean | Prior Std. Deviation |
|------------------------|------------|----------------------|
| Age | -100 | 625 |
| BOVAG Guarantee | 0 | 100 |
| Fuel (Diesel) | 1000 | 62500 |
| Fuel (Petrol) | -1000 | 62500 |
| Guarantee Period | 1000 | 62500 |
| HP | 50 | 100 |
| KM | -0.0625 | 0.125 |
| Manufacturer Guarantee | 1000 | 62500 |
| Quarterly Tax | 100 | 625 |
| Weight | 10 | 0.125 |

Results

For our final model, we decided to use model 4. This model is constructed using horse-shoe selected predictors with the scaled data and default priors. The horseshoe selected predictors include `Age`, `Mfg_Month`, `Mfg_Year`, `KM`, `Fuel_Type`, `HP`, `Metallic`, `QuartTax`, `Weight`, `Guarantee`, `BOVAG`, and `Period`. Also, using the scaled numeric predictors and a log-transformation of car price to restrict prediction prices to positive values only, this model is able to better predict selling price compared to other models.

```
## stan_glm
##   family:      gaussian [identity]
##   formula:     Price ~ Age + Mfg_Month + Mfg_Year + KM + Fuel_Type + HP + Metallic +
##                 QuartTax + Weight + Guarantee + BOVAG + Period
##   observations: 694
##   predictors:   28
##   -----
##                  Median MAD_SD
## (Intercept)    9.169  0.643
## Age          -0.120  0.379
## Mfg_Month2    0.022  0.028
## Mfg_Month3    0.033  0.051
## Mfg_Month4    0.012  0.075
## Mfg_Month5    0.034  0.097
## Mfg_Month6    0.017  0.122
## Mfg_Month7   -0.003  0.146
## Mfg_Month8    0.014  0.169
## Mfg_Month9   -0.039  0.194
## Mfg_Month10   -0.027  0.218
## Mfg_Month11   -0.030  0.242
## Mfg_Month12   -0.002  0.267
## Mfg_Year2000   -0.001  0.289
## Mfg_Year2001   -0.013  0.577
## Mfg_Year2002    0.100  0.862
## Mfg_Year2003    0.131  1.153
## Mfg_Year2004    0.102  1.447
## KM           -0.078  0.007
## Fuel_TypeDiesel 0.030  0.034
## Fuel_TypePetrol 0.171  0.035
## HP            0.050  0.006
```

```

## Metallic1      0.022  0.009
## QuartTax       0.076  0.009
## Weight         0.040  0.008
## Guarantee1    0.030  0.009
## BOVAG1         0.050  0.015
## Period          0.015  0.005
##
## Auxiliary parameter(s):
##     Median MAD_SD
## sigma 0.102  0.003
##
## -----
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg

## [1] "R^2:  0.881"

## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.

## [1] "LOO R^2:  0.87"

```

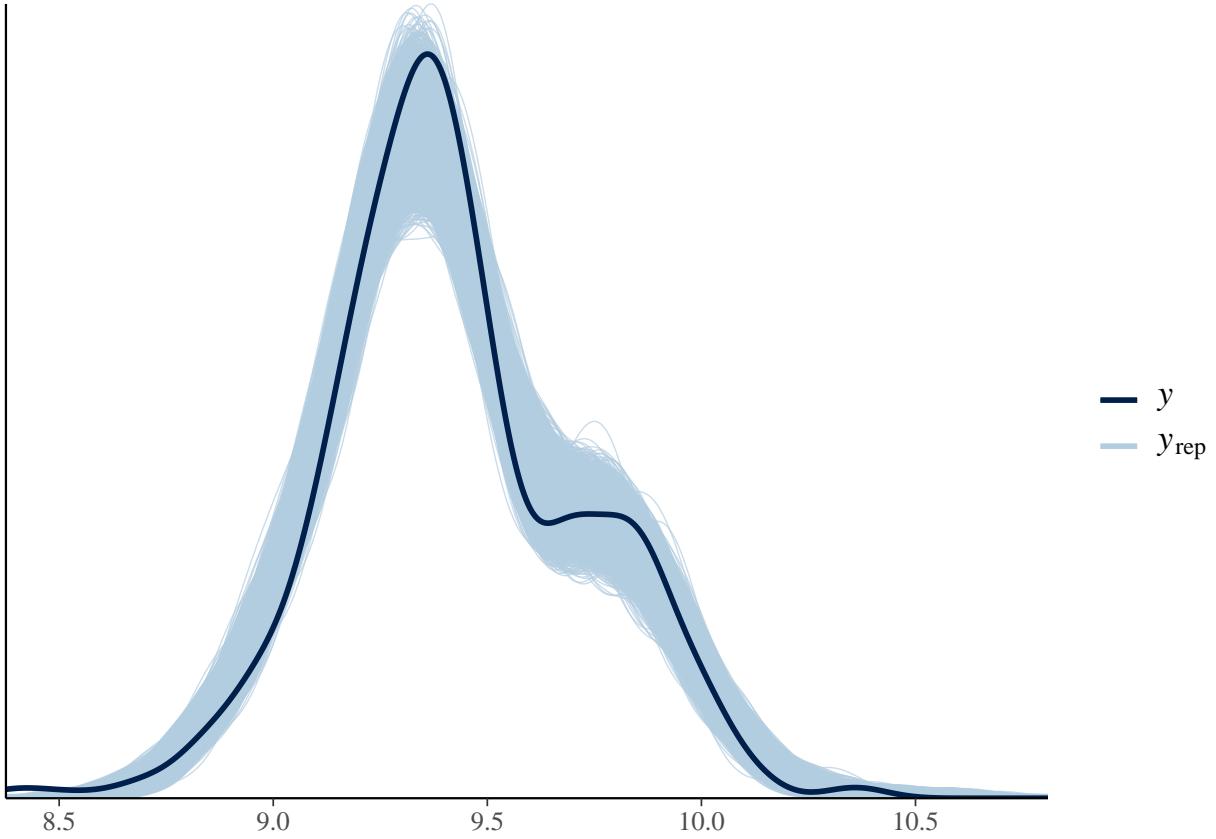
The intercept coefficient can be interpreted as the expected log(price) of a non-metallic, CNG-fueled car that was manufactured in January of 1999 with no manufacturer or BOVAG guarantee and a value of 0 for all numeric predictors included in the model. The coefficients estimates for the numeric predictors can be interpreted as the expected percent difference in predicted car price associated with a one standard deviation increase in the value of the predictor, with all other predictors remaining constant. Overall, this model had an in-sample Bayesian R^2 value of 0.882 and a leave one out adjusted R^2 value of 0.87. This confirms our model isn't overfitting while also showing a relatively high R^2 value. ADD SOMETHING ABOUT SIGMA!!!

The final fitted model should look like our data. By drawing from the predictive distribution and comparing it to the distribution of the response variable, we can determine if our model is fitting appropriately. In the plot below, we see our predictive distribution tracks the response distribution well increasing our confidence in the model.

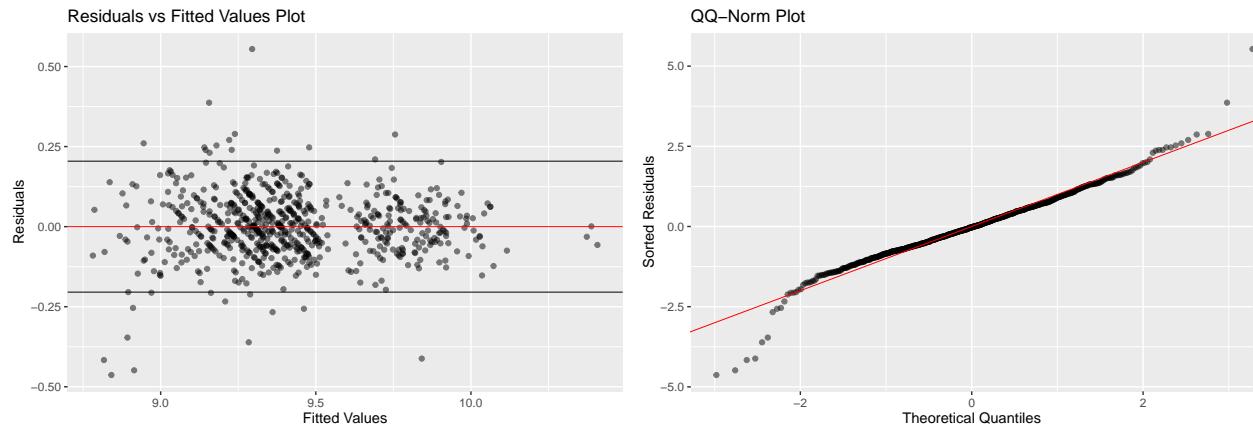
```

fit4_rep = posterior_predict(fit4)
ppc_dens_overlay(data_scale$Price, fit4_rep) + scale_y_continuous(breaks=NULL)

```



Checking the assumptions of our final model, we find the model includes all the relevant predictors as we chose these using the horseshoe prior and forcing predictors to be highly related with the response. The outcome measure accurately reflects our prediction interest and is generalized to all Toyota Corollas. Looking at the residual plots, we see there are no patterns and trends within the residuals vs fitted values plot. Most values are within 2 standard residuals of 0 and the values are spread across the 0 line. There might be some clumping but nothing big enough to question the model. There aren't any heavy tails in the qq-norm plot and the values closely align with the red line.



Others results (as appropriate)

Because the goal of this study is prediction, our final model has great observed predictive powers using leave one out cross validation and also k-fold cross validation. Using

In all, the purpose of this study was to predict the selling price of used Toyota Corollas and ensure a small profit based on their new purchase and trade-in promotion. Based on several variables, we were able to fit this model to help the dealership closely estimate the final selling price for their used cars. With this model, the dealership can now ensure a reasonable profit by plugging in the characteristics of each individual car into this model and output a predicted selling price. This will result in more accurate selling prices and higher profits for the dealer.