ECUTBILDNING

Statistical inference on used car data using linear regression in R

Linus Rundberg Streuli

2024-04-10

Table of contents

Abstract

This is were I put my abstract.

1 Introduction

1.1 Tasks and research questions

- Fit a regression model to predict the price of a used car given certain features
  - Collect data on used cars for sale
- Which features, if any, contribute to the predicted price?
  - Using statistical methods and domain research, select and possibly transform the most significant features in the data set
- Using cross validation, fit and evaluate a number of models and select the best performing model.
- Create a simple application to enter some data about a car and predict the price. (If there is time.)

2 Theory

2.1 Statistical learning

In short, statistical learning is the process of using statistical methods and mathematical models to understand data (James et al., 2023, p. 1). This process can be used to predict values given a set of variables, or to gain a better understanding of how those variables relate to the wanted value, also called *inference*. Often, the goal might be both prediction and inference.

Given that we wish to predict the response value Y from one or more variables, or predictors, X, this can be expressed in a general form as Equation 2.1.

$$Y = f(X) + \epsilon \tag{2.1}$$

Here, f is some kind of transformation made on X in order for it to help us estimate Y, end ϵ is an *error term* which is unknown, and assumed to be normally distributed with a mean of 0. We need ϵ because our model will never be able to take into account for every little variable that might influence or prediction.

One of the simplest, yet most powerful, methods of statistical learning is *linear regression*.

2.2 Linear regression

2.2.1 Simple linear regression

At the heart of linear regression lies the equation of the straight line (Equation 2.2)

$$y = b + mx \tag{2.2}$$

where b denotes the *intercept* where the line crosses the y-axis, and m denotes the *slope* of the line.

In simple linear regression we assume that the relationship between the response variable Y and the independent variable X is approximately linear, which can be expressed as (Equation 2.3). (James et al., 2023, p. 61)

$$Y \approx \beta_0 + \beta_1 X \tag{2.3}$$

Here, β_0 denotes the intercept and β_1 denotes the slope of the line. Using training data, we can estimate the *coefficients* $\hat{\beta}_0$ and $\hat{\beta}_1$ and predict a certain value of Y by computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{2.4}$$

Estimating coefficients

The coefficients can be estimated using a number of approaches, but the most common method is the *least squares* criterion.

The goal is to find values for β_0 and β_1 such that the difference between the observed response value and our predicted response value, the *residual*, is as small as possible. The ith residual e_i can be expressed as $e_i = y_i - \hat{y}_i$. Given n observations in our training data, the *residual sum of squares* (RSS) is defined as

$$RSS = e_1^2 + e_2^2 + ... + e_n^2, \tag{2.5}$$

or

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{2.6}$$

The least squares method then chooses β_0 and β_1 to minimize the RSS.

Evaluating model performance

To assess how well a model fits the data, two measures are commonly used: the *residual standard error* (RSE), and the R^2 statistic.

It is rarely the case that a models captures every thinkable variable that affects the outcome. Therefore, the error term ϵ from Equation 2.1 is present in our model. Equation 2.3 can be subsistuted for the general function f, which gives us Equation 2.7:

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{2.7}$$

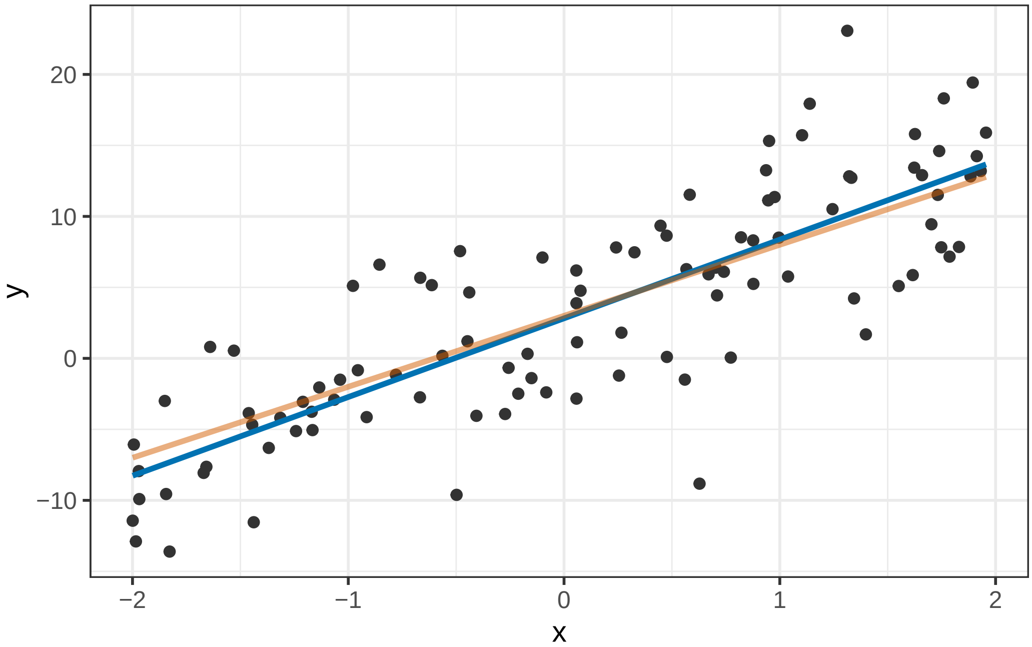


Figure 2.1: Linear regression on simulated data. The orange line shows the true relationship $f(X) = 3 + 5X$, and the blue line is the least squares estimate for f(X) based on the generated data.

The residual standard error

The RSE is an estimate of the standard deviation of ϵ (James et al., 2023, p. 69) and is computed by

$$RSE = \sqrt{\frac{1}{n-2}RSS}. \tag{2.8}$$

If the predicted values of the model are close to the observed values, the RSE will be small and the model fits the data well.

The RSE can also be used to find out if there is a relationship between the predictor and the outcome.

We can perform a hypothesis test with $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ to see if the coefficient value is sufficiently far from 0. By calculating the t-statistic

$$t = \frac{\beta_1 - 0}{SE(\hat{\beta}_1)} \tag{2.9}$$

we can find the probability of finding this value in the t-distribution. This probability is the p-value, and a small enough p-value, based on our chosen confidence level, might lead us to reject the null hypothesis that there is no association between the predictor and the response.

The R^2 statistic

The RSE is measured in the units of Y. This makes interpreting what a good value of the RSE is hard. (James et al., 2023, p. 70). The R^2 statistic, on the other hand, is a *proportion*, and as such always takes on a value between 0 and 1.

The R^2 statistic describes how much of the total variance in the response Y that can be explained by using X as a predictor. The formula used is seen in Equation 2.10.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.10}$$

$TSS = \sum (y_i - \bar{y})$ is the *total sum of squares* and RSS is defined in Equation 2.6. While R^2 is independent of the units of Y, it can still be hard to determine what constitutes a good value of R^2. This depends on the problem at hand as different domains handle different kinds of data, with different properties and relationships.

2.2.2 Multiple linear regression

So far we have been using a single predictor X_1 to predict the response Y. In *multiple linear regression*, additional predictors are added to form the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{2.11}$$

where X_j represents the jth predictor, and β_j represents the effect of X_j on Y, or, in other words, the *average* effect on Y on a one unit increase in X_j, given that *all other predictors remain fixed.*

The coefficients $\beta_0, \beta_1, ..., \beta_p$ are estimated in the same way as in simple linear regression, using the least squares method.

The F-statistic

To find out if any of our predictors actually affect the response variable, we can perform a *hypothesis test.* Our null hypothesis H_0 is that none of the predictors are related to the outcome:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

which we test against the alternative hypothesis H_a

$$H_a : \text{at least one } \beta_j \text{ is non-zero.}$$

To perform the test, we compute the *F-statistic*,

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1).} \tag{2.12}$$

If H_0 is true, the expected value of the F-statistic is close to 1. If the alternative hypothesis is true, we expect F to be greater than 1.

Variable selection

The F-statistic tells us if at least one of the predictors are associated with the response variable, but it cannot tell us which this or these are.

We could look at the p-values of the individual variables, but if we have a lot of predictors, the probability of finding p-values below 0.05 by chance increases.

This is again taking too long. I might need to move forward.

Potential problems

In *Introduction to Statistical Learning*, James and co-authors list six of the most common potential problems when fitting linear regression models (James et al., 2023, p. 93). These are:

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

2.3 Evaluating models

The theory behind evaluating model performance. Cross validation.

2.3.1 Train/test data

The reason behind splitting data.

2.3.2 Evaluation metrics

Statistical methods, RMSE...

2.4 Feature selection

Significance. Statistical methods vs. domain knowledge. (Hyndman, 2011)

3 Method

3.1 Data collection

The data behind this analysis was scraped from the swedish site Blocket which is a marketplace for individuals and businesses.

3.1.1 First data filtering step

The data scraped was filtered by a number of parameters set beforehand:

- The price range was between 20 000 and 500 000 SEK
- No cars from before the year 2000
- Only ads from individuals, no businesses
- No commercial vehicles

The price range was chosen to obtain as much data as possible without having to deal with values too extreme. Figure 3.1 shows the distribution of car prices at the time when filtered for the other parameters above. The selected range of 20 000 to 500 000 SEK seems like a good compromise.

The year 2000 was chosen somewhat arbitrarily. The goal was to have a mix of newer and older cars in the data set, but not too old.

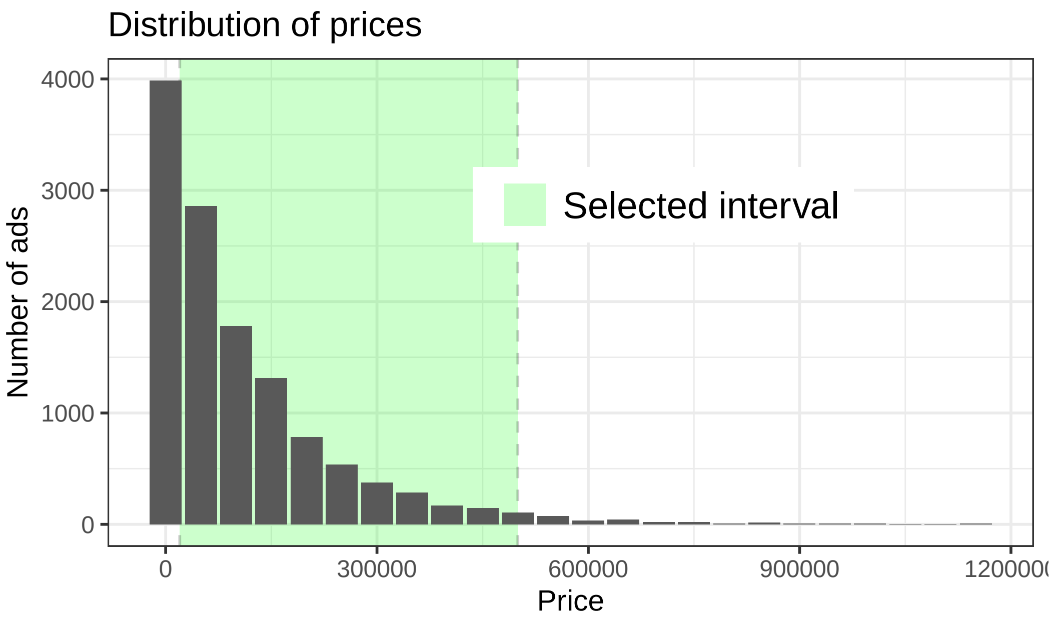
3.2 Data preprocessing

The data is imported and preprocessed in a pipeline. The comments in the code should be explanatory.

```
# Import data

df <- read.csv("data/car_ads_data_02.csv")

# Data preparation chain:
# Remove rows with price == NA
# Filter out prices < 20 000 and > 500 000
# Trim extra whitespace
# Convert character NA:s to actual NA:s
```



Source: blocket.se

Figure 3.1: Distribution of car ad prices

```
# Filter out rows with NA:s in type column
# Remove thousand separator whitespace in mileage
# Create make_model column and place it before fuel column
# Change column types
# Create age column
# Filter out cars with wrong first_in_traffic date
# Create days_in_traffic and miles_per_day columns
# Convert to tibble

df_prep <- df |>
  dplyr::filter(!is.na(price)) |>
  dplyr::filter(price >= 20000 & price <= 500000) |>
  mutate(across(everything(), str_trim)) |>
  mutate(across(everything(), ~ na_if(.x, "NA"))) |>
  dplyr::filter(!is.na(type)) |>
  mutate(mileage = str_replace_all(mileage, "\\D", "")) |>
  unite(make_model, make:model, sep= " ", remove = FALSE) |>
  relocate(make_model, .before = fuel) |>
  mutate(across(c(model, make_model, fuel, gearbox, type, drive, color, region), as.factor
  mutate(across(c(year, mileage, hp, motorsize, price), as.integer)) |>
  mutate(first_in_traffic = as.Date(first_in_traffic)) |>
  mutate(age = year(Sys.Date())-year) |>
  dplyr::filter(year(first_in_traffic) >= 1999) |>
  mutate(days_in_traffic = as.integer(Sys.Date() - first_in_traffic), miles_per_day = mile
  as_tibble()

# Change make to "Other" for observations < features
df_prep <- rows_update(
    df_prep, df_prep |>
    group_by(make) |>
        mutate(count = n()) |>
        dplyr::filter(count < length(df_prep)) |>
        mutate(make = "Other") |>
        select(-count),
    by = "id")

df_prep <- df_prep |> mutate(make = as.factor(make))
```

3.2.1 Train/test

Then the data is split into a training set and a test set.

```
# Split data
data_split <- initial_split(df_prep, prop = .8)

train_data <- training(data_split)
test_data <- testing(data_split)
```

The EDA section (-Section 4.1) of the Results chapter explains the data in more detail.

3.3 Training and evaluating models

For training and evaluting the models, various packages from the `tidymodels` framework were used. `rsample` was used to split the data into training and test sets, and for creating folds for cross validation. `recipes` and `workflows` were used to create a reusable pipeline for preprocessing the data. `parsnip` was used to create the actual linear regression model.

- How the data was prepared for fitting.
- Mark, Model and Mark/Model?
- Motor size does not make sense for electric cars. We need to decide whether to drop the column, or drop the electric cars. Start by looking at a model with the electric cars dropped to see if the motor size is significant for the remaining observations.

3.3.1 Variable selection

An interesting dilemma - can we create subsets of the mark predictors? Some of them are not significant while others are, but they describe the same aspect. Can we really train a final model using only a subset of marks? How will that affect predictions?

4 Results and discussion

4.1 EDA

The raw data was preprocessed as described in the Data Preprocessing section of the Theory chapter (-Section 3.2). Here we take a closer look at the data.

4.1.1 Overview

The entire data set consists of 7 661 observations with 18 features each. Each observation is an ad for a used car.

4.1.2 Missing data

Several of the observations in the raw data had missing values in the `price` column. Those where filtered out at the very start as `price` is our target value.

There are also some 700 observations that are have missing values in the `motorsize` column. Most of these are electric cars (see Figure 4.1). I want to keep the electric cars in the data set so I don't want to remove those observations all together.

There does seem to be some sort of relationship between motor size and price. (See Figure 4.2). Removing the column might have a negative impact on the model's ability to predict prices, but keeping it leaves me with a whole lot of missing values, and issues with electric cars. One way might be to impute the column with the mean or median of the motor sizes, but since electric cars actually don't have a motor size, that seems wrong.

My final decision was to remove the `motorsize` column.

4.1.3 Other columns with missing data

Table 4.2 shows the three remaining columns that have issing values. They are all categorical values and it doesn't make sense to impute them with the most common values. Since there will be a maximum loss of 123 observations, the rows with the missing values will be removed.

Missing values

Table 4.2: Columns with missing values

```
# A tibble: 7,535 x 18
   id       make  model make_model fuel  gearbox mileage  year type  drive    hp
   <chr>    <fct> <fct> <fct>      <fct> <fct>     <int> <int> <fct> <fct> <int>
 1 1400504~ Audi  A3    Audi A3    Bens~ Manuell  207611  2005 Halv~ Fyrh~   200
 2 1400489~ Skoda Octa~ Skoda Oct~ Dies~ Manuell   21218  2014 Kombi Fyrh~   105
 3 1400612~ BMW   420   BMW 420    Dies~ Automat   19586  2015 Halv~ Tvåh~   184
 4 1400546~ Kia   Opti~ Kia Optima Dies~ Automat   23700  2012 Sedan Tvåh~   136
 5 1400603~ Ford  S-Max Ford S-Max Dies~ Automat   20850  2012 Fami~ Tvåh~   200
 6 1400631~ Volk~ Golf  Volkswage~ Bens~ Automat   10840  2017 Kombi Fyrh~   180
 7 1400643~ Saab  9-3   Saab 9-3   Bens~ Automat   26340  2006 Kombi Tvåh~   176
 8 1400617~ Merc~ E     Mercedes-~ Dies~ Automat   55423  2006 Kombi Tvåh~   150
 9 1400675~ BMW   525   BMW 525    Bens~ Automat   28557  2006 Kombi Tvåh~   218
10 1400592~ Audi  A6    Audi A6    Dies~ Automat   24219  2013 Kombi Fyrh~   204
# i 7,525 more rows
# i 7 more variables: color <fct>, first_in_traffic <date>, region <fct>,
#   price <int>, age <dbl>, days_in_traffic <int>, miles_per_day <dbl>
```

Table 4.1: Overview of training data

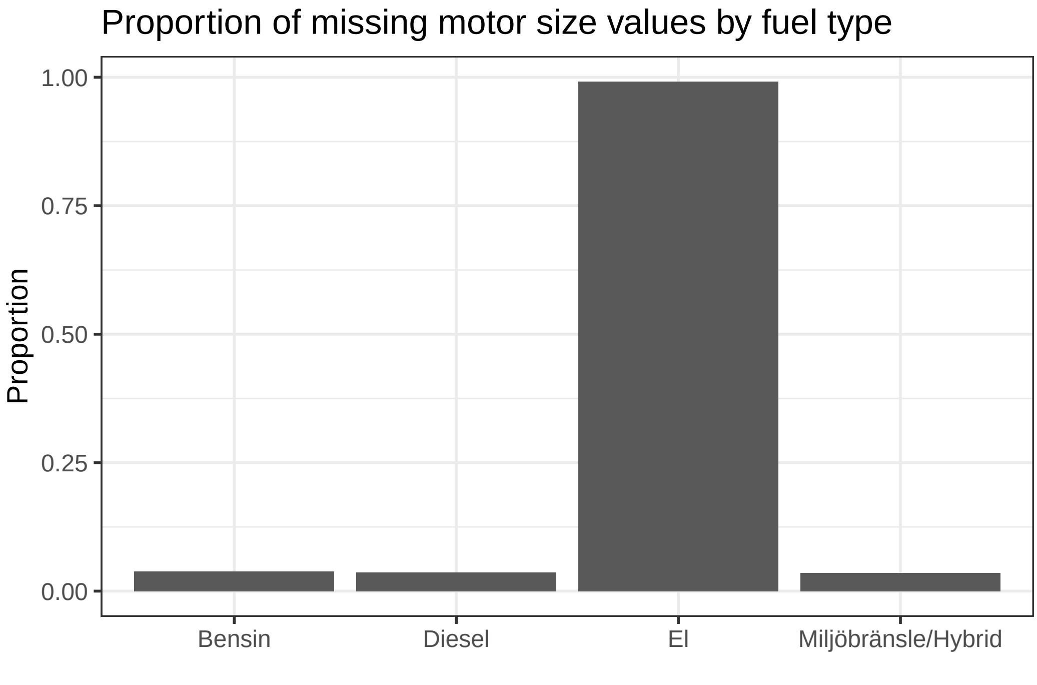


Figure 4.1: Proportion of missing motor size values by fuel type

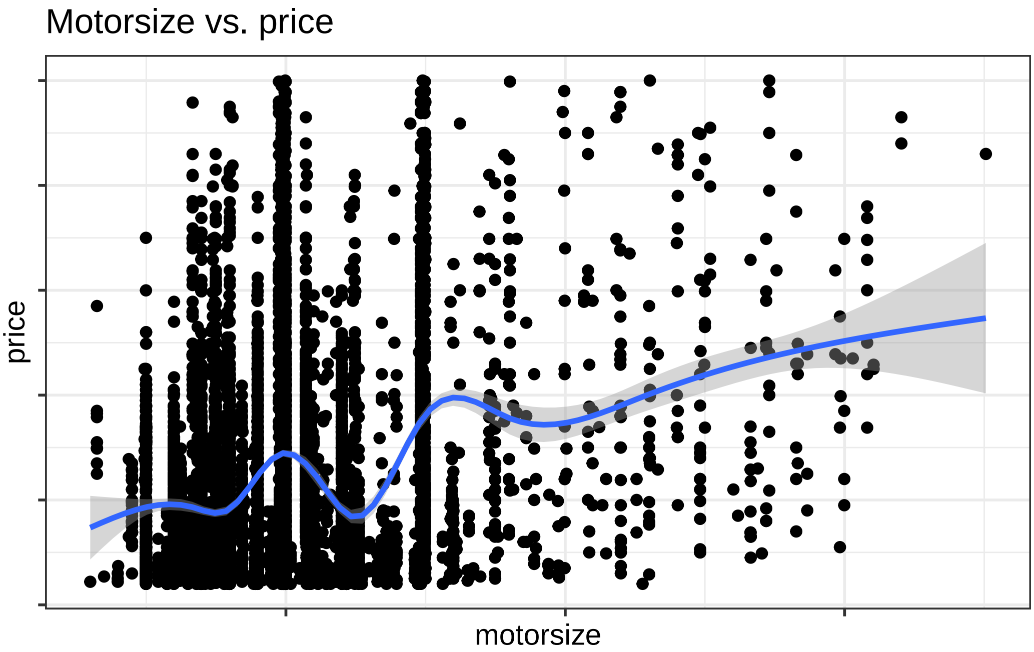


Figure 4.2: Motor size vs. price with an added regression line.

Some values in the `make` category were only present in a few observations. This increases the risk of having values that are only present in the training or test data sets. At least one make was present only in the single observation. Even when using a stratified split, that observation would end up in either the training or the test set.

I decided to keep them in the data, but group them together under the "Other" name. The cutoff is set to be less than the number of columns in the data set, that is, 18.

4.2 Results

4.2.1 Which car makes have a significant impact on the price?

Of the 63 unique car makes in the original data set, 28 had 18 observations or fewer. As mentioned, these are grouped together under the "Other" name, which gives us 36 unique values for the `make` category. One question I wanted to ask was, which makes (if any) seem to have an impact on the asking price for the car? Figure 4.3 shows the estimated coefficients of the 20 car makes that have a statistically significant effect on the price.

Any other car make has an average price close enough to the mean price that it can't be said that the make alone has any effect on the price.

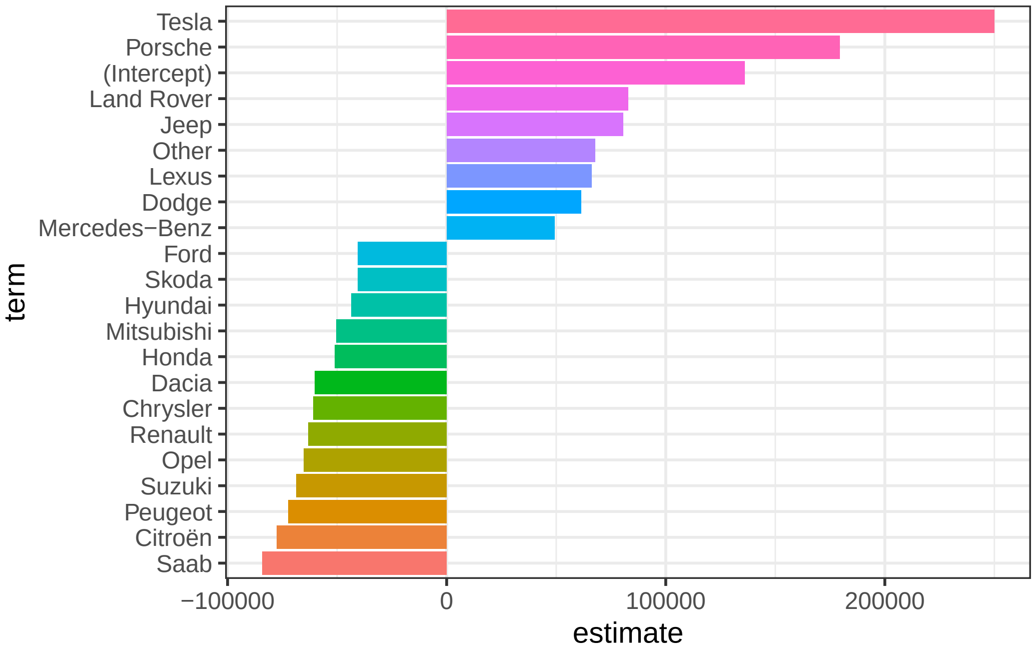


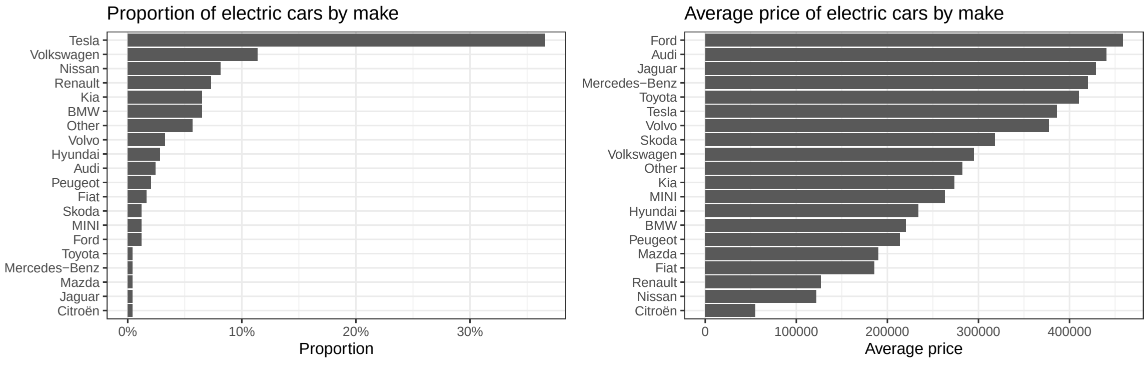
Figure 4.3: Estimated coefficients of car makes with a p-value < 0.05

4.2.2 Tesla is driving up the prices of electric cars

From Figure 4.3 we see that the car make that has the highest effect on price is Tesla.

Since Tesla exclusively makes electric cars, the question is how this might affect a model's ability to predict the prices of that kind of vehicle.

A vast majority



(a) Proportion of electric cars by make

(b) Average prices of electric cars by make

Figure 4.4: Proportions and average prices of electric cars by make

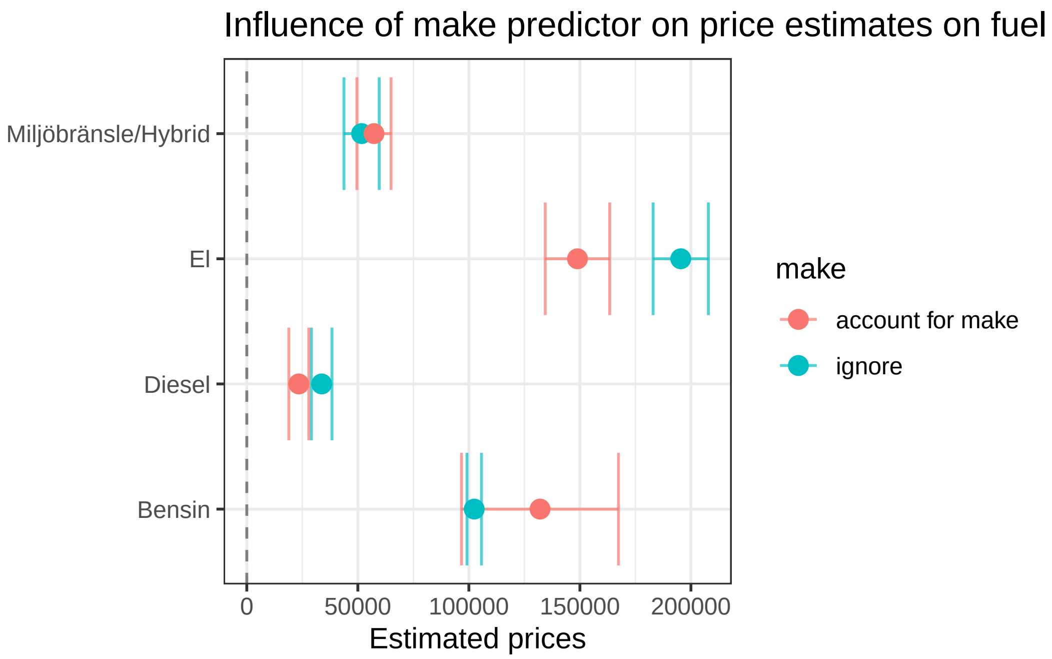


Figure 4.5: Confidence levels of coefficients for the fuel category, with and without taking the make category into account.

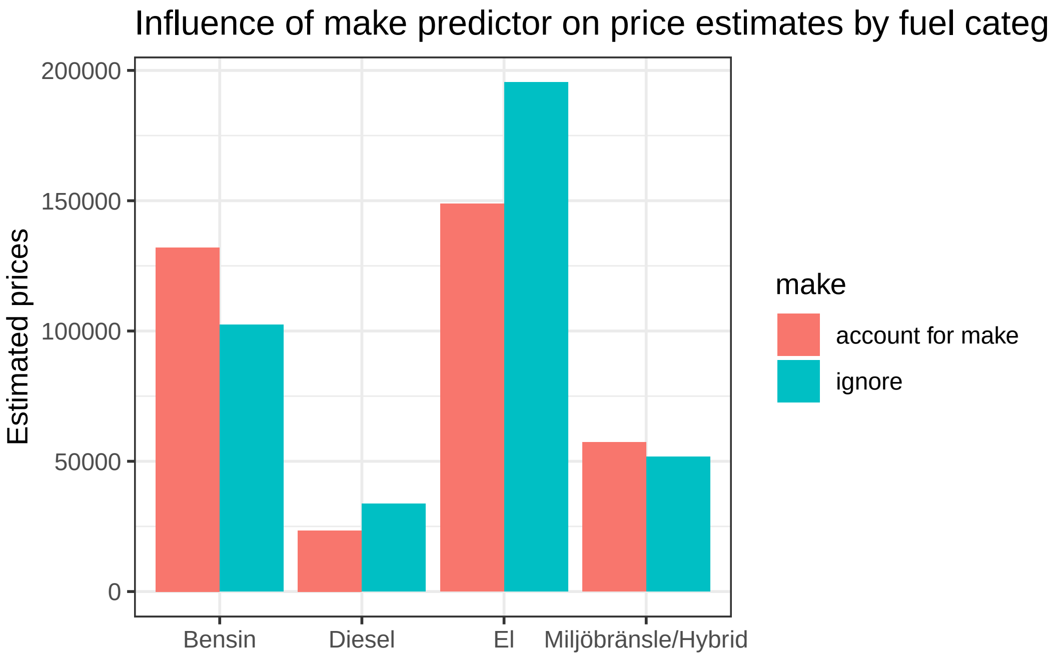


Figure 4.6: Influence of make predictor on price estimates by fuel category

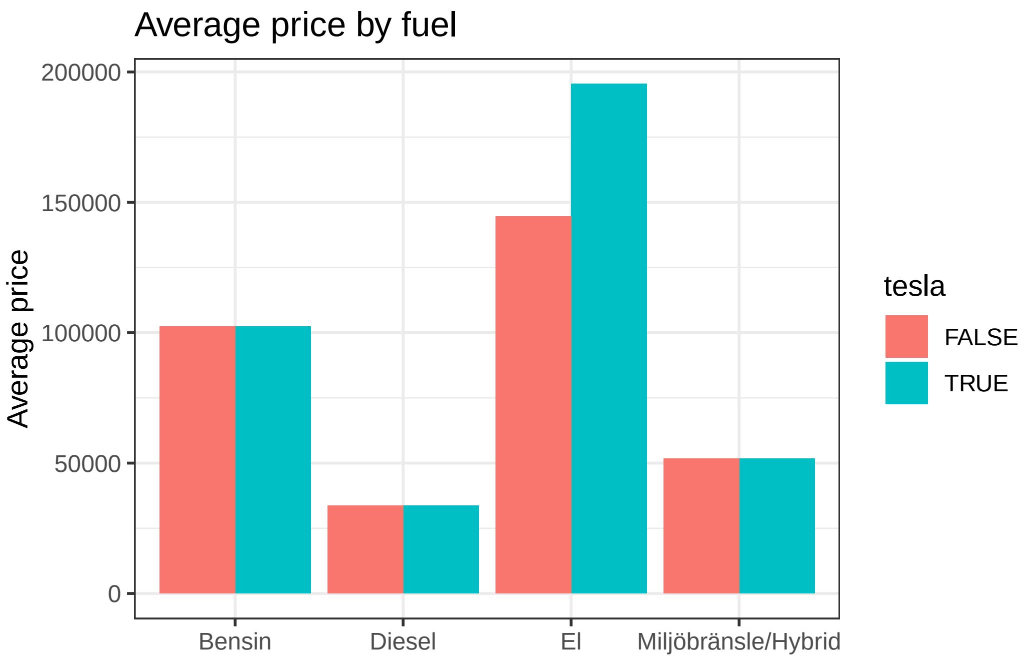


Figure 4.7: Average price by fuel type, with and without Teslas in the model.

4.3 Discussion

4.3.1 Legal and ethical issues regarding data mining

The data behind this analysis was collected using an automated method, often called *data mining* or *scraping*.

Blockets EULA (End User License Agreement) states the following regarding scraping their data (blocket.se, 2024):

> Du har inte rätt att kopiera, reproducera, publicera, ladda upp, skicka eller distribuera något material eller någon information från Webbplatsen utan föregående skriftligt tillstånd från Blocket. (...)
>
> Användning av automatiserade tjänster såsom robotar, spindlar, indexering eller liknande, samt andra metoder för systematisk användning av innehållet på Webbplatsen är inte tillåtet utan föregående skriftligt tillstånd från Blocket.
>
> All otillåten användning medför ersättningsskyldighet. Den som avsiktligt eller genom grov oaktsamhet bryter mot lagen kan straffas med böter eller fängelse upp till två år och bli dömd att betala skadestånd.

Or, in english, (my translation and emphasis):

> You are not allowed to copy, reproduce, publish, upload, send or distribute any material or any information from the web site without prior written consent from Blocket. (...)
>
> Use of automated services such as robots, spiders, indexing or the like, and other methods for systematic use of the web site's content is prohibited with prior written consent from Blocket.
>
> Any prohibited use comes with an obligation to compensate. Anyone who, knowingly or by gross negligence, *breaks the law* can be punished by fine or prison for up to two years, and be ordered to pay damages.

It is not clear from the text which law is referred to. The EULA also states that Blocket owns the immaterial rights to any material such as text, images, design and information meda available by using the site. This, then would be a question of copyright law. It is however not immediately clear that the contents scraped from the web site is such that it would fall under copyright law. [1]

Sweden is however a member of the European Union, and in 1996 the European Council approved the *Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases*, (European Union, 2019).[2]

The directive, amended in 2019, works as an analog to the copyright laws, and protects the rights of database creators and owners. This might be the law that is referred to in the Blocket EULA.

Should probably write Blocket and ask them!

- Difference b/w copyrighted material such as images, and information in a database. Also difference b/w US and European law.
- EU *sui generis* law protects database owners but exceptions could be made for research. European Union (2019)

4.3.2 Data analysis

- Would new car price be a significant variable? We can only guess as we didn't include it.
- Gathering the description data and performing textual analysis might give more insights. The car might be broken down, deregistered, or any other number of factors that affect the asking price but do not show up in the data.
- Collect the same data over time to find patterns
- Electric cars are underrepresented - what will that to to the data? <- Introduction?

[1] citation needed

[2] Check how it actually works

4.3.3 Parameter selection

Statistical methods vs. domain knowledge

5 Conclusions

6 Summary

References

blocket.se. (2024). *Användarvillkor*. https://www.blocket.se/om/villkor/anvandarvillkor

European Union. (2019). *Directive 96/9/EC of the european parliament and of the council of 11 march 1996 on the legal protection of databases*. https://eur-lex.europa.eu/eli/dir/1996/9/oj/eng

Hyndman, R. J. (2011). *Statistical tests for variable selection*. https://robjhyndman.com/hyndsight/tests2/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An introduction to statistical learning: With applications in r* (2nd ed.). Stringer New York, NY.