Predicting used car prices with 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

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

2.2.1.1 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.

2.2.1.2 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:

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

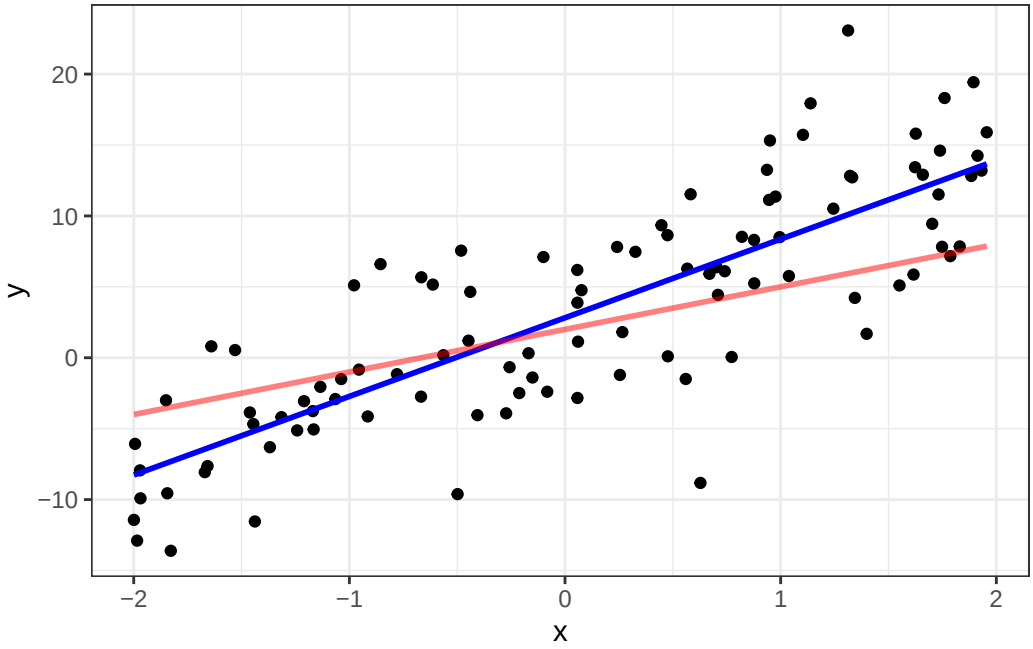

Figure 2.1: Linear regression on simulated data. The red line shows the true relationship $f(X) = 3 + 5X$, and the blue line is the least squares estimate for f(X) based on the generated data.

2.2.1.2.1 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.

2.2.1.2.2 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 2.9.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.9}$$

$TSS = \sum (y_i - \bar{y})$ 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

2.2.2.1 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

How we collected the data and why we selected the parameters that we did. Did not collect price for new car.

3.2 Training and evaluating models

- 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.

3.2.1 Parameter selection

4 Results and discussion

4.1 EDA

4.1.1 Considerations regarding data

4.2 Results

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.
- Collect the same data over time to find patterns
- Electric cars are underrepresented - what will that to to the data? <- Introduction?

4.3.3 Parameter selection

Statistical methods vs. domain knowledge

[1] citation needed
[2] Check how it actually works

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.