A linear regression analysis on car data

Filip Holmberg

EC Utbildning

Kunskapskontroll R

202404

Abstract

In the work for the current report data from sales ads of cars was collected and used to create a linear regression model. The regression analysis did not indicate violation to any associated assumption, which support the results of inference. For numeric variables, the price decrease with -1,8 % per 1000 Swedish mile and increases with 6.6 % per model year (with a level of confidence of 95 %). Categorical variables such as fuel type, gearbox and brand will also affect the selling price. The model's prediction accuracy made its practical uses limited, which was believed to be a result of generally high variation of car prices on the market.

List of abbreviations and concepts

Swedish	English
Pris	Price
Bränsle	Fuel type
Växellåda	Gearbox
Miltal	Milage
Modellår	model year
Biltyp	Car type
Drivning	Drive
Hästkrafter	Horsepower
Märke	Brand
Halvkombi	Hatchback
Kombi	Station wagon
El	Electricity
Miljöbränsle	Bio-Diesel or Ethanol
Bensin	Gasoline
Tvåhjulsdriven	Two wheel drive
Fyrhjulsdriven	Four wheel drive

Table of Contents

1 Introduction

Our daily lives involve many activities which are geographically scattered, and transportation is fundamental for our way of life,

“The relationships between transportation and society are numerous, deep, varied, ancient, and complex. Any summary of them sounds trite. Everyone has had extensive personal experiences using transportation. Transportation has influenced each of our choices about where to live, spend vacations, shop, or work. So inescapable is the tie between transportation and society that, like gravity, we take it for granted and cannot imagine a world without it.” (Kulash, 2000)

Luckily in Sweden the infrastructure often support transportation in cities with many options: by bus, train, car, cycling, walking and others (i.e. electrical scooters). The further away from major municipalities, the more restricted these options become. In the outskirts sometimes the only available option is going by car. It can be agreed by many that the car will continue to be an import mean of transportation. In fact, in the period between 2014-2023 the amount of newly registered cars was as high as 300742 in average per year (SCB, 2024).

Data driven insight have become more and more important in the modern automotive industry. Being able to draw conclusions (inference) and make predictions can help stakeholders make the best of their car deals. The linear regression model is a useful tool, as it allows us to identify relationships between variables and predict outcomes, entirely based on observations in historical data.

The main purpose of this report is to analyze the market value of cars in Sweden. To achieve this objective the following task and research questions are relevant:

1. Create a linear regression model for car prices from sales ads.
1. Which variables affect the price of a car (inference)?
2. How accurate are the predictions of price made by the model?

2 Theory

2.1 Multiple linear regression models for dependent relationship

Suppose that there is a relationship between the dependent variable Y and independent random variables X_i. A multiple linear regression model can be described as follows,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \qquad \text{(eq. 1)}$$

here ε is a random variable which denotes the residual. It corresponds to the deviation from the regression line. β_0 is intercept to the function and the coefficient β_i is the derivative with respect to X_i, with the others held constant. The intercept and coefficients parameters are unknown constants and are estimated from the data (inference).

2.1.1 Statistical significance in multiple linear regression models

The statistical significance of random variables in multiple linear regression models can be judged based on the following test of hypothesis,

$$H_0: \beta_0 = \beta_1 = \ldots = \beta_n = 0 \qquad \text{(eq. 2)}$$

and

$$H_1: at\ least\ one\ \beta_i \neq 0 \qquad \text{(eq. 3)}$$

It is common to adapt to a level of confidence of 5 %. In the case of one variable (i equal to 0 and 1) the p-test is used, where p < 5 % is considered significant (H_0 is rejected). However, there is an inherent chance of 5 % that the variable obtains a value discarding the null hypothesis (false positive). The more variables that the model contain the more likely the event that a parameter is faulty considered significant. The F-test can be used regardless to number of variables and is used similarly to the p-value.

2.1.2 Confidence and prediction interval

In the case of car sales, the mean of the population can be used to describe the market value of a car, at a certain level of X. The equation for a regression line, for a population, can described as (S. Körner, 2006),

$$Y = \beta_0 + \beta_1 X \qquad \text{(eq. 4)}$$

where β_0 and β_1 are unknown constants. At a certain level of X, the corresponding unknown and constant Y lies within the confidence interval (CI). That is to say, the true mean, or *market value*, of the population lies within this interval (with a e.g. 95 % level of confidence).

The prediction interval (PI) is broader than that of the CI. This is due to the increased variance coupled to the residual (see eq. 1). The PI correspond to a prediction of an individual, rather than a mean of the population (CI). In other words, the predicted Y of a certain individual at a particular level of X lies within the PI (with a e.g. 95 % level of confidence).

2.2 Selecting variables using best subset and forward stepwise selection procedures

Selection of variables can be done in various ways. The aim is often to maximize performance of model on test error while keeping the model "simple" (e.g. few predictors). In this report best subset selection and forward stepwise selection procedures were considered for selection of variables (James et al., 2023).

The best subset selection method is as accordingly,

1. The sample mean is predicted for each observation, i.e. the null model.
2. For k = 1,2, …, p: $\binom{p}{k}$ models containing k predictors are fitted. The model with the highest Adjusted R^2 is considered best.
3. Chose between models in step 1 and 2 using the prediction error (i.e. BIC or Adjusted R^2).

The forward stepwise selection procedure is similar to that of best subset selection. The difference is in step 2:

2. For k = 0, …, p-1: Augment the previous predictor(s) using p-k models. The best among these p-k models are chosen on the basis of highest Adjusted R^2.

2.3 Training, validating and testing models

Splitting data into sets of training, validation and test is a common approach for development, evaluating respectively testing models. In this approach the selection of variables and training the model is made on the training set. The validation set is used for evaluation of models. Finally, the test set is used for the sake of probing how the model would perform on new data.

This approach using these 3 datasets is not necessary in modelling of linear regression, as some measures (R^2_{adj} and BIC) indirectly estimates the test error (see section 2.4). However, we still adapt to the 3 dataset approach due to the following:

- The approach is straight forward and devoid data leakage.
- The number of data entries are large (> 1500 observations). Hence there is not obvious need of data conservation.
- It simplifies the reuse of code for studying the sample size effect on confidence interval.

2.4 Measures for model evaluation

Measures are used to summarize certain aspects of models. The measures used in the work for this report are shown in Table 1.

Table 1: Measures for evaluating models.

Measure	**Measure (equation)**	**Description**
Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ (eq. 5)	Mean error of predictions on the validation/test data.
Residual Sum of Squares	$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$ (eq. 6)	Measures the unexplained variation.
Total Sum of Squares	$TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ (eq. 7)	Measures the total variance.
Determination coefficient	$R^2 = 1 - \frac{RSS}{TSS}$ (eq. 8)	Measures the fraction of variance of Y which is explained by the relationship with the independent variables X_i.
Adjusted R^2	$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$ (eq. 9)	Estimate test error from training error. It compensates increase of R^2 when adding more variables by a decrease (penalty). Parameters n and p represent sample size respectively number of independent variables. The aim is to maximize this measure.
Bayesian Information Criterion	$BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$ (eq. 10)	Estimate test error from training error. Parameters n, d and $\hat{\sigma}^2$ present sample size, number of predictors and estimate of variance of error, respectively. The smaller the BIC value, the better.

2.5 Potential problems for uses of a linear regression model

The most common potential problems that arise when fitting a linear regression model to a dataset is described in Table 2 (S. Körner, 2006), (James et al., 2023). Keeping track of these aspects help the model describe the data better. Therefore, more reliable inference and prediction conclusions can be drawn.

Table 2: Potential problems and effects when fitting a linear regression model. Suggestions on evaluation and troubleshooting is also briefly described.

Potential problem	Potential effects	Evaluation	Suggested troubleshooting
1. Non-linearity of response-predictor relationship	Model does not represent data well at all. Unreliable predictions and inference.	Plot residuals vs fitted values.	Include non-linear transformations of the random variable(s).
2. Correlation of residuals	CI, PI and p-values become underestimated.	Plot residuals vs variable [Svante, p382], residuals should not form a pattern.	Reconsider approach in collection of data.
3. Non-constant variance of residuals (heteroscedasticity)	CI, PI and hypothesis test rely on homoscedasticity.	Scale-location plot.	Transform dependent variable, i.e. log(Y).
4. Non-normal distribution of residuals	CI, PI and hypothesis test assumes normal distribution of residuals.	QQ-plot and histogram of residuals.	See troubleshooting in row 1, 3, 5.
5. Outliers	CI, PI and p-values.	Studentized residuals: should be within -3 to 3.	Deemed outliers should be removed/make a better model.
6. High-leverage points	Affect regression line.	Residual vs leverage plot or studentized residuals vs leverage plot.	Consider benefit of removal of points vs limitations of model.
7. Multicollinearity	Standard deviation of β_i increases. Rejecting false H_0 becomes more difficult.	Variance Inflation Factor (VIF) should be below 5.	Variables with high multicollinearity: remove/combine.

2.5.1 Diagnostic plots for linear regression analysis

There are 4 main diagnostic plots for linear regression analysis (B. Kim, 2015). These are shown in Figure 1 and briefly described in the following list:

- Residual vs fitted. Non-linearity of response-predictor relationship is evaluated. Residuals should not form a pattern (e.g. parabola shape) and preferably randomly distributed at 0.
- Quantile-quantile (QQ) plot. Ideally the standardized residuals should follow a straight line. Deviations indicate that residuals are not normally distributed.
- Scale-location plot. The assumption of homoscedasticity is evaluated. The square root of standardized residuals should be randomly spread.
- Residuals vs leverage. Points that are potentially influential on the regression results show up outside the dashed lines. These points have high Cook's distance scores, > 0.5-1.

An additional diagnostic plot that addresses leverage high-leverage points and outliers is the Studentized residual vs leverage plot. This plot can be seen in Figure 2. When the absolute value of studentized residual exceeds 3, the datapoint is considered a potential outlier. Potential high-leverage points are found far from any other datapoint in the x-axis of the plot. Here, point 41 could potentially be a high-leverage point and point 20 a potential outlier.

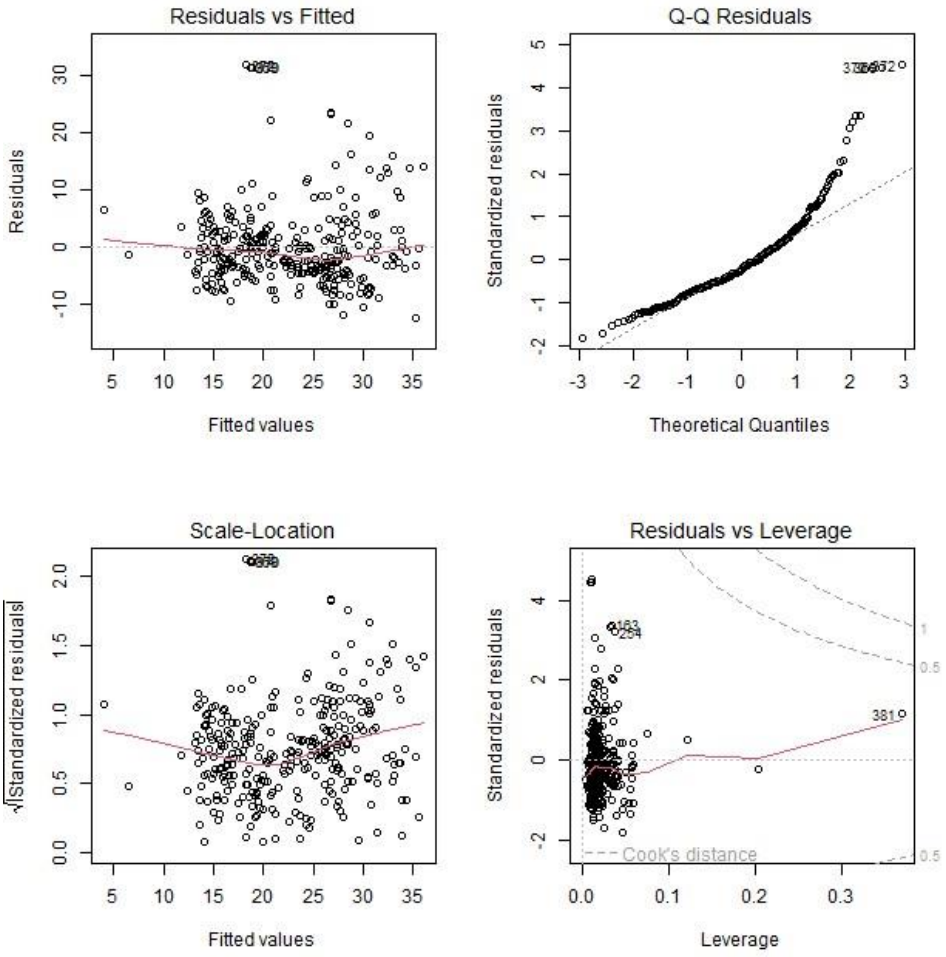

Figure 1: Four diagnostic plots showing Residuals vs Fitted (top left), QQ-plot (top right), Scale-location (bottom-left) and Residual vs leverage (bottom-right). Boston data has been modelled for the sake of producing these plots (Newman et al., 1998).

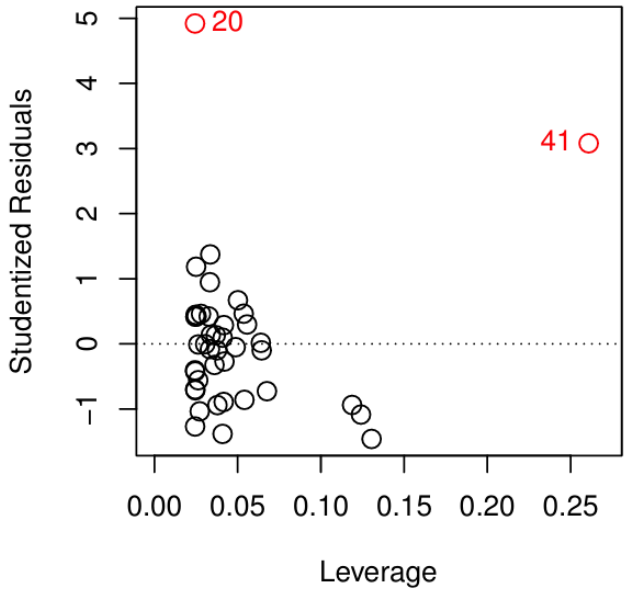

Figure 2: Plot showing studentized residals vs leverage (James et al., 2023, s. 98).

3 Method

The work for the current report was made entirely in R. The code is available elsewhere (F. Holmberg, 2024).

3.1 Data: collection, cleaning and splitting

The following list describes the workflow of the data before regression analysis,

- A data frame was constructed by collecting data from sales ads at external source (Blocket, 2024). The dependent and random variables are shown in Figure 3. A total of 2726 observations were collected. The search filter was set to 150-500 k SEK and did not allow the cartype “family van” or “commercial vehicles”. Only a few popular brands were considered (Volvo, Renault, Volkswagen, Toyota, Ford and Audi).
- Data was cleaned by removing duplicates and rows containing missing values. “Price”, “milage”, “model year” and “horsepower” were formatted to the numeric class, while the rest were set to factorial class. The column “color” was removed as it was discovered to contain incorrect values. After cleaning a total of 2576 observations remained.
- The dataset was randomized and split into the sets of training (60 %), validation (20 %) and test (20 %).

	Pris	Bränsle	Växellåda	Miltal	Modellår	Biltyp	Drivning	Hästkrafter	Färg	Märke
1	229900	miljöbränsle/hybrid	automat	1994	2022	halvkombi	tvåhjulsdriven	116	vit	toyota
2	269900	miljöbränsle/hybrid	automat	0	2023	halvkombi	tvåhjulsdriven	117	grå	toyota
3	189900	bensin	automat	7609	2017	suv	fyrhjulsdriven	116	grå	toyota
4	419900	miljöbränsle/hybrid	automat	6500	2021	suv	fyrhjulsdriven	306	vit	toyota
5	254000	miljöbränsle/hybrid	automat	689	2021	halvkombi	tvåhjulsdriven	123	vit	toyota
6	304900	miljöbränsle/hybrid	automat	1651	2022	suv	fyrhjulsdriven	117	gul	toyota

Figure 3: The 6 first rows of the data frame (R) showing the dependent variable "Pris" and random variables. Labels are given in Swedish.

3.2 Workflow for regression analysis

Unless otherwise stated, the training set was used for the below steps of the workflow,

1. Brief data exploration for the sake of overviewing the data.
2. Create initial model (including all parameters) and test the hypothesis that the value of a car depends on at least one of the variables.
3. Evaluate which variables to include by study the effect of variable selection on Adjusted R^2. Here, best subset and forward stepwise selection approach was used.
4. Evaluate model(s) on the validation set. The need of a validation set is only due to the use of the RMSE measure. Here, we use the validation set in order to keep track of improvements made in the following steps 4 and 5. If model of satisfaction proceed to next step.
5. Diagnose potential problems (see section 2.5). Adjustments are made and model is evaluated (step 4). If model is of satisfactory, proceed to step 6.
6.
    a. Evaluate which parameters affect the car price (inference).
    b. Evaluate prediction accurary on the test set.

4 Results and discussion

4.1 Data exploration

A summary of the training set is shown Figure 4. Overall, this summary represents a snapshot of the sales ads available at one point in time. In total there are 8 random variables, which constitutes 17 predictors after encoding.

```
      Pris                    Bränsle      Växellåda       Miltal        Modellår        Biltyp             Drivning    Hästkrafter          Märke
 Min.   :150000   bensin            :532   automat:1329   Min.   :    0   Min.   :2014   cab     :  8   fyrhjulsdriven:659   Min.   : 73.0   audi      :542
 1st Qu.:199800   diesel            :584   manuell: 216   1st Qu.: 3420   1st Qu.:2017   coupé   : 23   tvåhjulsdriven:886   1st Qu.:142.0   ford      :181
 Median :249800   el                :105                  Median : 7200   Median :2020   halvkombi:351                        Median :191.0   renault   :196
 Mean   :272152   miljöbränsle/hybrid:324                  Mean   : 7833   Mean   :2019   kombi   :537                        Mean   :187.8   toyota    :206
 3rd Qu.:329000                                            3rd Qu.:11760   3rd Qu.:2021   sedan   : 74                        3rd Qu.:207.0   volkswagen:182
 Max.   :499900                                            Max.   :24896   Max.   :2024   suv     :552                        Max.   :561.0   volvo     :238
```

Figure 4: Printout from R of summary of training dataset.

4.2 Hypothesis testing of independent relationships

The summary of a linear regression model that includes all parameters is shown in Figure 5. As the F-statistic is low (< 2.2e-16) it can be concluded that the response “Price” has a dependence on at least one variable. It can also be seen that many variables have low (< 0.05) p-values, which is consistent with this conclusion.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -3.413e+07  1.220e+06 -27.986  < 2e-16 ***
Bränslediesel              2.728e+04  3.087e+03   8.837  < 2e-16 ***
Bränsleel                  3.208e+04  4.537e+03   7.072 2.33e-12 ***
Bränslemiljöbränsle/hybrid -1.772e+02  3.434e+03  -0.052 0.958843
Växellådamanuell          -2.609e+04  3.187e+03  -8.186 5.63e-16 ***
Miltal                    -5.120e+00  3.134e-01 -16.337  < 2e-16 ***
Modellår                   1.702e+04  6.040e+02  28.186  < 2e-16 ***
Biltypcoupé               -3.453e+04  1.589e+04  -2.173 0.029951 *
Biltyphalvkombi           -7.811e+04  1.397e+04  -5.591 2.68e-08 ***
Biltypkombi               -7.265e+04  1.405e+04  -5.172 2.62e-07 ***
Biltypsedan               -5.884e+04  1.451e+04  -4.056 5.24e-05 ***
Biltypsuv                 -5.157e+04  1.401e+04  -3.682 0.000239 ***
Drivningtvåhjulsdriven    -1.642e+04  2.890e+03  -5.683 1.58e-08 ***
Hästkrafter                7.725e+02  2.103e+01  36.729  < 2e-16 ***
Märkeford                 -2.314e+04  3.706e+03  -6.244 5.50e-10 ***
Märkerenault              -4.453e+04  3.839e+03 -11.598  < 2e-16 ***
Märketoyota               -2.341e+03  4.162e+03  -0.562 0.573952
Märkevolkswagen           -1.976e+04  3.466e+03  -5.700 1.43e-08 ***
Märkevolvo                 3.011e+03  3.122e+03   0.964 0.335022
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38430 on 1526 degrees of freedom
Multiple R-squared:  0.8208,    Adjusted R-squared:  0.8187
F-statistic: 388.4 on 18 and 1526 DF,  p-value: < 2.2e-16
```

Figure 5: Summary of a linear regression model containing all 17 variables.

4.3 Selection of variables using best subset and forward stepwise selection

The procedure of best subset and forward stepwise selection resulted in much similar model behavior, as seen in Figure 6. The rate of increase in Adjusted R^2 decreases as the number of predictors increases. Somewhere around 10-15 predictors the rate of increase appears to cease. At this point the aim of the model needs to be considered. In general, that would be (1) reduce the estimated test error (Adjusted R^2) and (2) simplify the model (reduce number of predictors). Fewer variables reduces risk for overfitting (Géron, 2019). For the sake of completeness, it should also (3) keep/remove all predictors from each respective variable (i.e. "Märke"). Thus, one viable compromise is the removal of the variable "Drivning" (with encoded predictors "fyrhjulsdriven" and "tvåhjulsdriven").

With little adjustment (removing the variable "Drivning") of the model the Adjusted R^2 model is high (0.81). This indicate that the model would perform well on new data. However, the model needs to be reviewed from regression analysis perspectives to evaluate assumptions made (see 4.4).

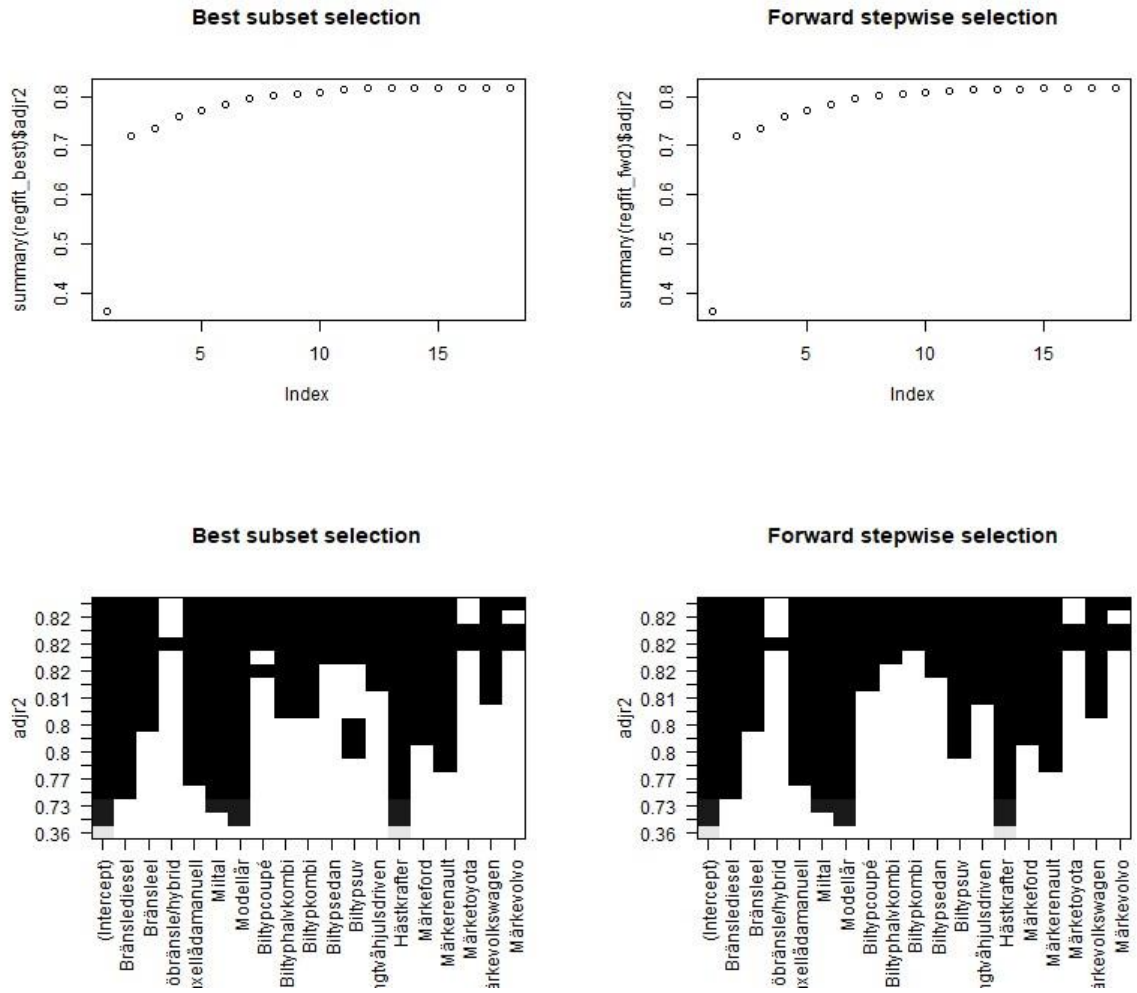

Figure 6: Adjusted R^2 against numbers of predictors (top) and contribution of Adjusted R^2 by predictor (bottom).

4.4 Evaluation of models on validation set

Several models containing different variables were considered. The performance of each model on the validation set was measured and is shown in Table 3. The results show that removal of the variable “Drivning” did affect any of the measures substantially. In addition, by transformation of the dependent variable “Pris”, a lowering of BIC was obtained. On the basis of these results, it deemed only worth to evaluate model 3 on its reliability in terms of inference and prediction.

Table 3: Measures of 3 different models on the validation set.

Model	Variables	Adjustments	RMSE	Adjusted R^2	BIC
1	All	None	39017	0,819	37132
2	All-“Drivning”	None	39667	0,815	37157
3	All-“Drivning	Log(“Pris”)	39731	0,825	-1741

4.5 Diagnosis of and adjustments to potential problems

The linear regression assumptions were evaluated for Model 3, and the resulting diagnostic is discussed in the list below:

1. The residual vs fitted values plot show no visual pattern (see Figure 7). This indicates the assumption of non-linearity has been met. However, there is a strange diagonal band appearing. Further inspection of the data reveal that each level of “Pris” form a diagonal line of dots. This is highlighted in Figure 8, and is entirely a result of collecting the data with the restriction of 150-500 k SEK. Since residuals within this span appear randomly distributed, there is no indication that beyond these restrictions they would be distributed differently. All in all, we believe the model to represent the data at a satisfactory level.
2. The assumption of non-correlation of residuals was tested and the results are seen in Figure 9. Only the variables of “Modellår” and “Hästkrafter” show feasible signs of correlated residuals. To be precise, one may suspect patterns in these two respective plots. For the “Hästkrafter” variable, there are too few observations around 500 to verify any potential pattern. Hence we discard the notion that the assumption is violated (by this variable). For the residuals coupled to “Modellår” a pattern resembling a wave can, possibly, be seen. We believe this wave-pattern to be too vague in order to verify it as correlated residuals.
3. There are no immediate signs of heteroscedasticity in the scale-location plot (see Figure 7), as residuals are randomly distributed. A small tendency of heteroscedasticity appears when fitted values are above 13.3, but overall it is indicated that the residuals are homoscedastic.
4. Overall, the QQ plot in Figure 7 indicate that residuals are normally distributed. This is verified by the histogram shown in Figure 10, which also appears to be normally distributed. A QQ plot of model 2 can be seen in Figure 11. For model 2 there is a tendency of higher deviation of the residual from the line. Therefore the transformation of “Pris” to log(“Pris”) improved the model in two ways: it lowered the BIC score and was also helpful in verifying the assumption of normal distribution of residuals.
5. In Figure 12 it is seen that more than few studentized residuals spread beyond the formal limit of suspected outliers (3 and -3). Therefore we adjust this formality to 4 and -4, which results in 3 potential outliers (see Figure 13). We can only speculate that these vehicles have attributes that stand out of the general population (i.e. imported cars or other). This could cause effect on listed prices, which would produce abnormal residuals. Since these points cannot be verified as outliers, they were allowed to remain in the training set. Since the sample size is large (n = 1545), the impact is likely to be small (R.W. Nahhas, 2024).
6. There were no single high-leverage points identified in the training set (see Figure 12). Instead, there are small cluster located at the leverage of 0.12-0.14. Inspecting the data reveal that these points all belong to the cartype “cabs”. Since these belong to a car type which would be interesting to model, they were not removed.
7. VIF value were calculated and the results can be seen in Table 4: VIF values for the 7 variables in model 3. All VIF values are below 5 and this verifies that there are no multicollinearity between variables.

To clarify and summarize the model’s limitations discussed above, there are no indications that the model would be practically limited in its practical applications (i.e. inference, predictions, etc.).

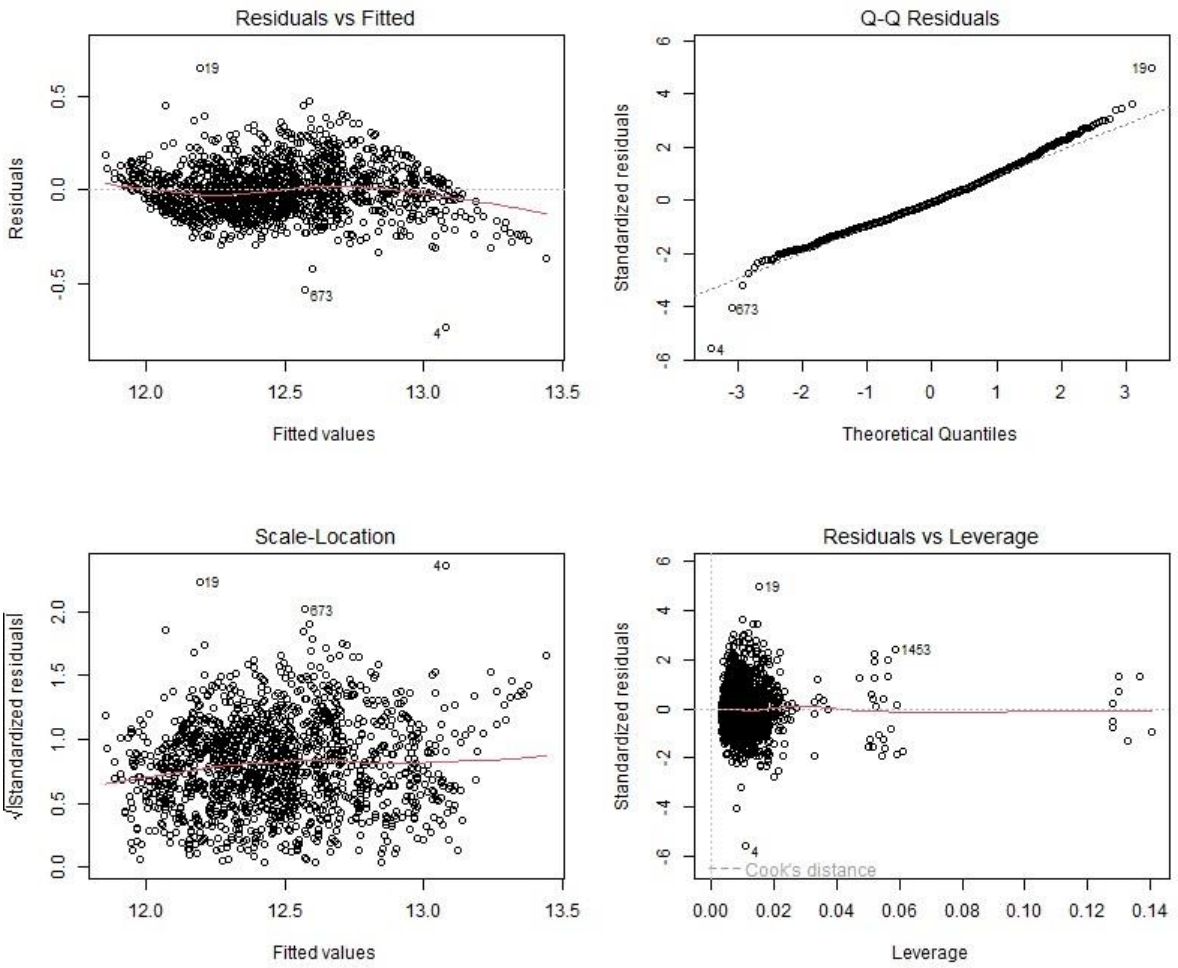

Figure 7: Diagnosis plots of model 3.

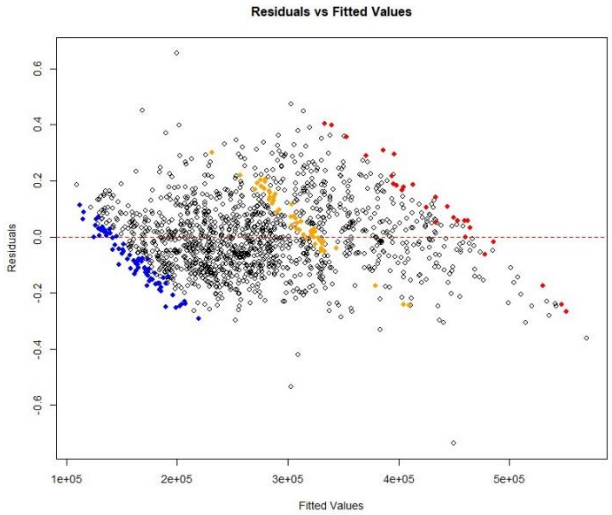

Figure 8: Residual vs fitted values plot highlighting cars sold at 150-160 k SEK (blue), 300-320 k SEK (orange) and 490-500 k SEK (red).

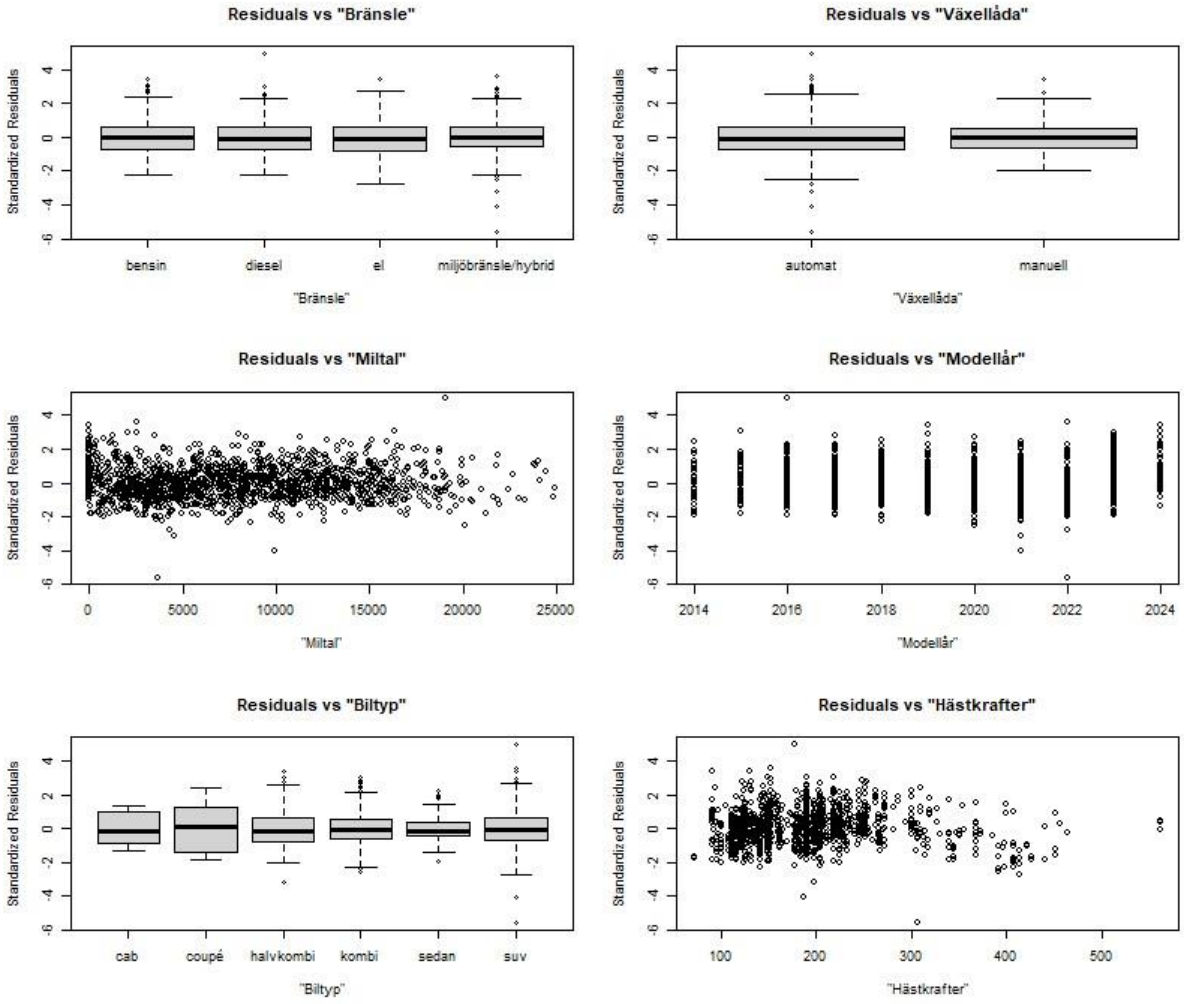

Figure 9: Box and scatter plots of residuals vs variables.

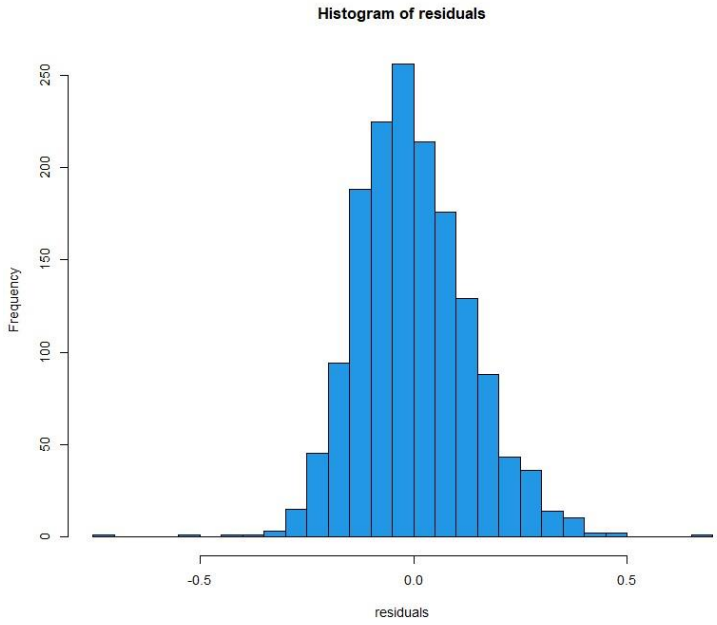

Figure 10: Histogram of residuals using model 3.

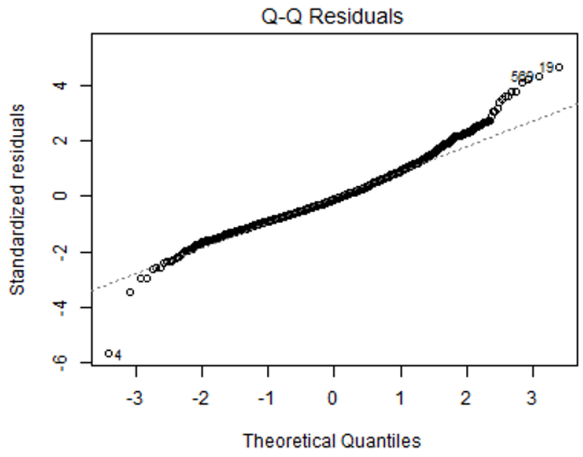

Figure 11: QQ plot of model 2.

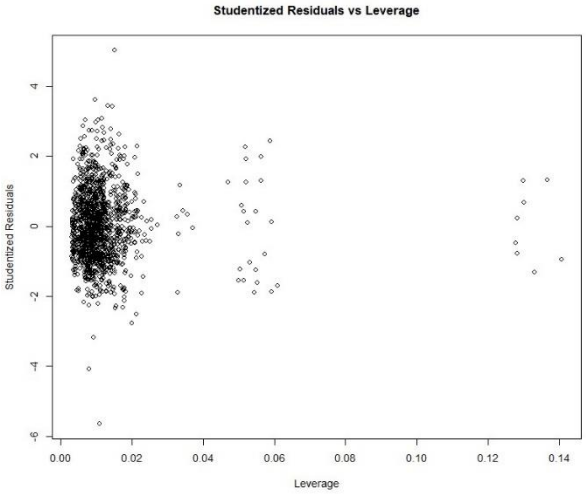

Figure 12: Studentized residuals vs leverage (model 3).

Pris	Bränsle	Växellåda	Miltal	Modellår	Biltyp	Drivning	Hästkrafter	Märke
229900	miljöbränsle/hybrid	automat	3692	2022	suv	fyrhjulsdriven	306	toyota
379900	diesel	automat	19097	2016	suv	fyrhjulsdriven	177	toyota
169000	miljöbränsle/hybrid	automat	9900	2021	suv	tvåhjulsdriven	187	toyota

Figure 13: Printout from R of potential outliers (studentized residual limit was set to 4 and -4).

Table 4: VIF values for the 7 variables in model 3.

Variable	**VIF**
Bränsle	3,5
Växellåda	1,3
Miltal	3,0
Modellår	2,6
Biltyp	1,7
Hästkrafter	1,6
Märke	2,8

4.6 Inference and predictions made by applying the model to test set

Inference made by the model show that predictors affect the price differently (see Table 5). To clarify, the intercept contain the following vehicle attributes: gasoline, automatic, cab, audi). Values presented are in comparison to this baseline. For instance, the fuel typ Diesel increases the price by 13 %, as compared to a vehicle which run on gasoline (keeping all other attributes similar). According to the model it can be concluded that the following attributes increase the price of a car most:

- Fuel type: Diesel.
- Gear box: Automatic.
- Car type: Cab. Surprisingly, the model suggest there is a great (15-25 %) price difference to other car types.
- Brand: Audi.

As stated previously, it is a design choice to keep "cabs" in the dataset. A consequence of this is the risk of altering the coefficients of the model. This could explain the large price difference between types of cars. Therefore, it would be wise to reconsider the design choice and exclude such cars from the model. Removing variables can reduce the risk of overfitting, making it better at generalize new data.

Meanwhile the numeric variables affect the price as expected,

- Increased milage decreases the price (-1,8 % per 1000 Swedish miles).
- The modelyear increase the price (6.6 % per "year").

We also note that the amount of horsepower did not alter the price much (> 1%). All in all, the inference made by the model is as expected for the two numerical variables "Miltal" and "Modellår". In addition, it also points out that the price is also affected by other factors ("Växellåda", "Brand").

Table 5: Confidence interval for coefficients in the linear regression model.

Predictor	**exp(Estimate)**	**2.5 %**	**97.5 %**	**Percent (%)**
(Intercept)	2,54217E-51	6,75E-55	9,57E-48	0
Bränslediesel	1,134	1,112	1,157	13,4
Bränsleel	1,045	1,014	1,078	4,5
Bränslemiljöbränsle/hybrid	0,994	0,971	1,017	-0,6
Växellådamanuell	0,880	0,862	0,900	-12,0
Miltal	0,999982	0,999980	0,999984	-0,0018
Modellår	1,066	1,062	1,070	6,6
Biltypcoupé	0,887	0,797	0,988	-11,3
Biltyphalvkombi	0,752	0,684	0,826	-24,8
Biltypkombi	0,764	0,695	0,840	-23,6
Biltypsedan	0,801	0,726	0,883	-19,9
Biltypsuv	0,831	0,757	0,914	-16,9
Hästkrafter	1,003	1,003	1,003	0,3
Märkeford	0,896	0,874	0,918	-10,4
Märkerenault	0,828	0,807	0,850	-17,2
Märketoyota	0,980	0,953	1,008	-2,0
Märkevolkswagen	0,930	0,908	0,952	-7,0
Märkevolvo	1,002	0,981	1,023	0,2

To understand the models accuracy in making prediction, it is necessary to evaluate the CI and PI of the model. In Figure 15, the histogram of 515 observations from the test set is displayed, for CI respectively PI. These histograms do not appear to be normally distributed. Hence the following discussion need to be more general.

We expect the PI to be much wider than CI, which is confirmed by comparing the ranges shown. This is due to the increased variance coupled to residuals. The CI can be interpreted as the interval in which the true mean of the market price lies. The actual price of an individual car should at 95 % of the time lie within the PI interval. From this we can argue to the following:

- Because the PI interval is very wide (around 50-300 kSEK), we therefore expect the difference between the real price and the point estimate (PI) to be spread (e.g. tails) out at equal proportions (25-150 kSEK).
- The CI is less wide (around 0-50 kSEK) and contain the true mean market value. Therefore we expect the distribution differences between the real price and the point estimate (CI) to be centered somewhere between -25 and 25 kSEK).

In Figure 14 the histogram of differences (actual price-predicted price) is shown. The mean of this distribution was calculated to 1094 SEK and tails appear at approximatively minus/plus 25-150 kSEK, respectively. This is in agreement with the above discussed expectations. Therefore, one can coupled the models accuracy (shown in the histogram) with its CI and PI. Therefore the accuracy depends on the same factor: variance coupled to residuals, namely too much noise in selling prices of cars.

Unfortunately, it is deemed that the model's predictability is too inaccurate for any practical uses. This conclusion is drawn with the notion that one simply does not wish to follow up on a predicted price and risk spending e.g. 100 kSEK more than needed.

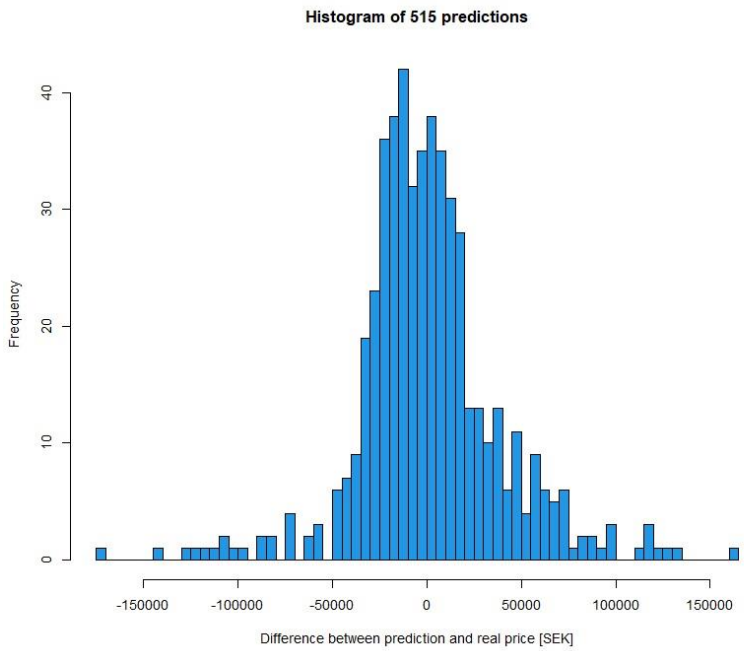

Figure 14: Histogram of 515 predictions using model 3.

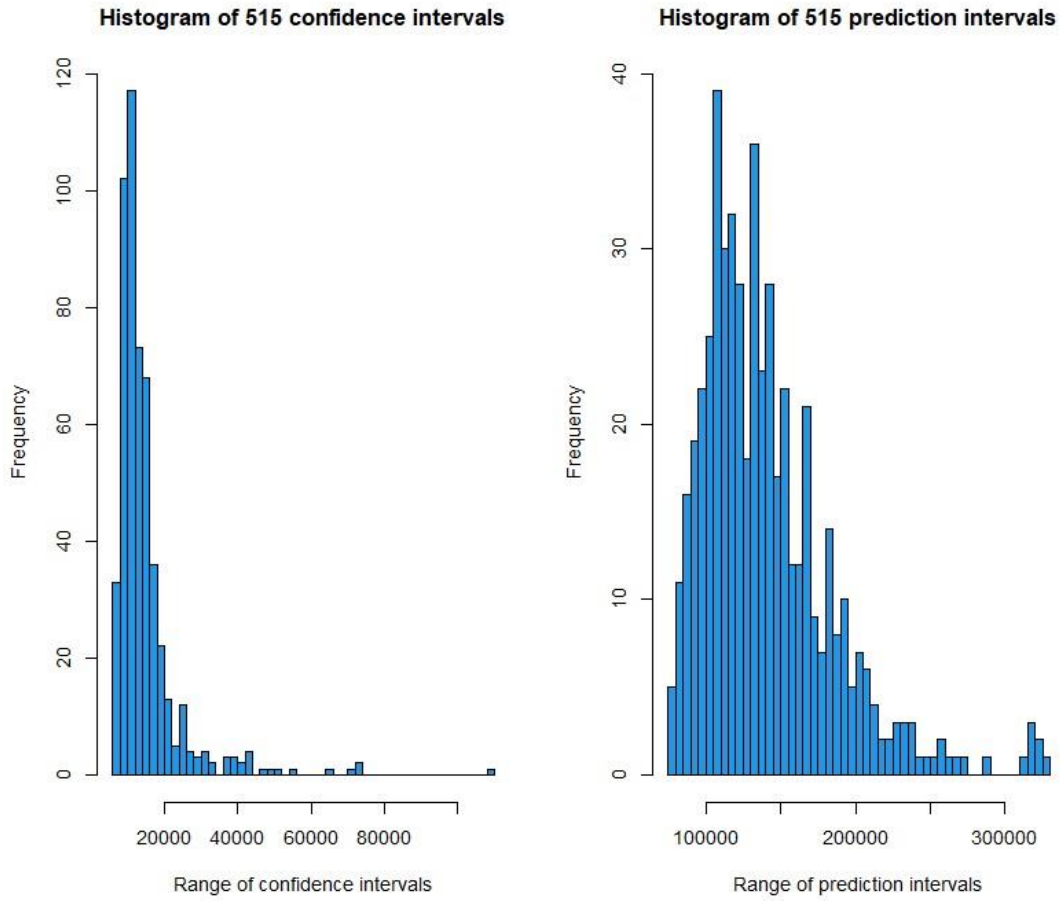

Figure 15: Histograms of confidence (left) and prediction (right) intervals using model 3 on test data.

5 Conclusions

The following conclusions regarding set out objectives and questions can be drawn,

1. A linear regression model for car prices was successfully created. Diagnosis of the model revealed no immediate violation to assumptions made. It was decided to include the car type “cabs” (high-leverage points) although it comes with the risk of affecting the models coefficients.
2. From the inference made with the model it was possible to conclude that:
   - Increased milage decreases the price with -1,8 % per 1000 Swedish miles.
   - The model year increase the price with 6.6 % per “year”.
   - Fuel type, gearbox and brand also affect prices. In addition, the model suggest that car type also influence the price.

   However, we believe the model would benefit by removing the car type variable, as it was not considered reasonable for the price to differ 15-25 % depending on this variable. Possibly, that could improve the model’s ability to generalize new data.

3. The accuracy of the model was evaluated, and its predictions was deemed not useful. Differences between selling prices and predictions were not normally distributed. Still, it was suggested that the mean of these differences should lie between -25 and 25 kSEK, which was verified experimentally (1 kSEK). However, the spread of the difference was 300 kSEK. Hence, the models practical use of predicting prices are limited. This was coupled to high variation in car prices, in general.

6 References

B. Kim. (den 21 09 2015). Hämtat från https://library.virginia.edu/data/articles/diagnostic-plots

Blocket. (den 20 04 2024). Hämtat från https://www.blocket.se/

F. Holmberg. (den 26 04 2024). Hämtat från https://github.com/FilipHolmbrg/R

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn Keras & TensorFlow.* Canada: O'Reilly.

James et al. (2023). *An Introduction to Statistical Learning.*

Kulash, D. (2000). Transportation and Society. i *TRANSPORTATION PLANNING HANDBOOK* (s. 4).

Newman et al. (1998). *http://www.ics.uci.edu/~mlearn/MLRepository.html*. Hämtat från http://www.ics.uci.edu/~mlearn/MLRepository.html

R.W. Nahhas. (2024). *www.bookdown.org*. Hämtat från https://www.bookdown.org/rwnahhas/RMPH/mlr-outliers.html

S. Körner, L. W. (2006). *Statistisk Dataanalys.*

SCB. (den 26 04 2024). Hämtat från www.scb.se