# What determines the price of a Volkswagen?

Andreas Rasmusson - EC Utbildning

April 25, 2025

**Abstract**

In this paper, we use statistical inference via linear regression to come to the following (not so shocking) conclusions:

1. Volkswagen cars with more horsepower tend to be more expensive.
2. Volkswagen cars with higher mileage tend to be less expensive.
3. Newer model years of Volkswagen cars tend to be associated with higher prices.
4. Manual transmissions are a reliable path to saving money (and stalling in traffic).

While none of these insights are worthy of the Nobel Prize, we take great pride in demonstrating that even the most intuitively true statements deserve the full force of statistical analysis. Because, sometimes, the obvious needs a confidence interval.

# Contents

# 1 Introduction

The number of cars in Sweden has, not surprisingly, seen an increase in the years 2011 - 2023, but the increase started to level out around 2020, as can be seen in figure 1.1 below:

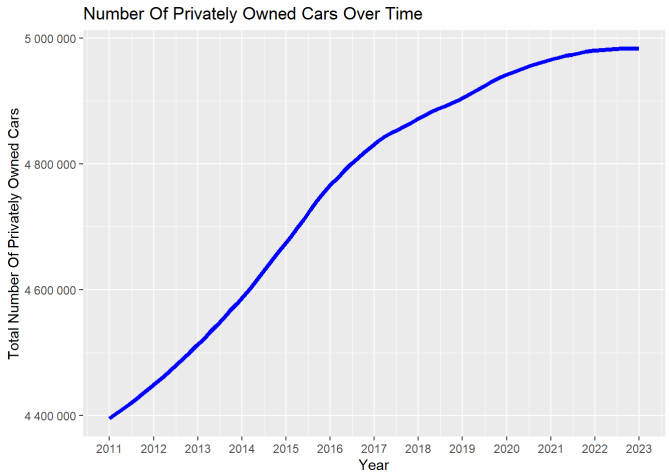


Figure 1.1: Number of privately owned cars in Sweden over time (Statistics Sweden (2025))

At the same time, perhaps because of Covid and geopolitical factors, the median yearly income has gone from increasing to decreasing over these years (Figure 1.2 below):

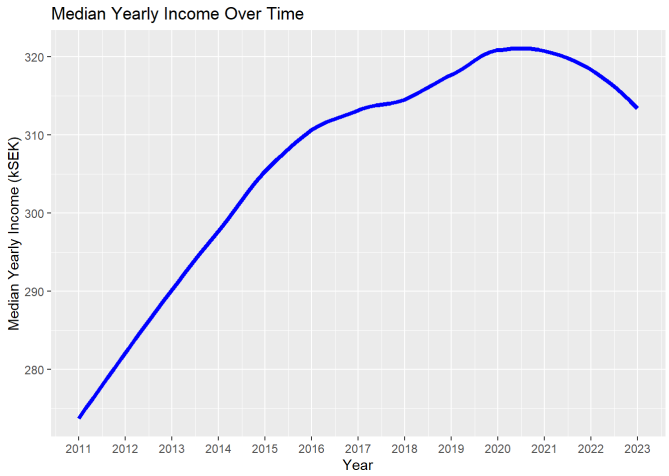


Figure 2.2: Median yearly income (kSEK) in Sweden over time (Statistics Sweden (2025))

This development suggests that perhaps people just can't afford to buy cars as much as they have in the past and the market and prices of cars may reflect that. In this paper, we examine what variables determine the price of a car. More precisely, we examine a dataset of Volkswagen car ads, containing the following variables:

- Försäljningspris (ask price)
- Säljare (seller - private or company)
- Bränsle (fuel)
- Växellåda (transmission)
- Miltal (mileage)
- Modellår (model year)
- Biltyp (type of car)
- Drivning (two-wheel or four-wheel drive)
- Hästkrafter (HK) (horsepower)
- Färg (color)
- Modell (model)
- Region (region)

The non-processed dataset contains approximately 1200 observations. Using this dataset, we ask and answer the following:

## 1.1 Research questions

1. Which sub collection, if any, of these variables significantly explain the variance in ask price for these ads?
2. Can we train a linear model that predicts unsean data at an $rmse/mean(ask\ price)$ of at most 0.2?
3. Seeing as a linear model is highly biased, do we see a dramatic improvement in $rmse/mean(ask\ price)$ if we instead train a highly complex model, such as an xgboost regressor on this data?

# 2 Theoretical background

## 2.1 Linear regression

In this sub section we introduce the linear regression model (James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021)). The linear regression model is a statistical model for estimating the relationship between a scalar response variable $Y$ and a set of explanatory variables $X_1, X_2, ..., X_m$. The model assumes that the relationship satisfies the following equation:

$$Y = \epsilon + \beta_0 + \sum_{k=1}^{m} \beta_k X_k$$

Here $\epsilon$ is assumed to be unknown but normally distributed. The $\beta$ parameters are the unknown quantities to be estimated from the training data. Training data is assumed to be an i.i.d sample of some size $n$ from the joint distribution of $X = (X_1, X_2, ..., X_m)$ and $Y$. Note that the "$X$-part" of the training data is a an $n \times m$ matrix

$$X_{train} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

and the "$Y$-part" is a column vector

$$Y_{train} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

We can estimate $Y$ by choosing some coefficient vector $\beta = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{pmatrix}$ and let the estimation $\hat{y_i}$ of Y, given an observation $(x_{i1}, x_{i2}, ..., x_{im})$ of $X$ be

$$\hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{m} \hat{\beta}_k x_{ik} \tag{1}$$

Note that, if we append a column vector of ones to the far left of $X_{train}$ to get what we call a design

matrix $X_d$, and let $\hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}$, we can write equation (1) in matrix form as follows:

$$\hat{Y} = X_d\hat{\beta}$$

How should the column vector $\hat{\beta}$ chosen? Since we have $n$ observations of $Y$ for $n$ observations of $X$, it would be a good idea to choose $\hat{\beta}$ in such a way that $X_d\hat{\beta}$ is as close to $Y_{train}$ as possible. How should this closeness be measured? Well, $X_d\hat{\beta}$ and $Y_{train}$ are both vectors, so we could simply use euclidean distance:

$$||X_d\hat{\beta} - Y_{train}|| = \sqrt{\sum_{i=1}^{n}(X_{d,i}\hat{\beta} - y_i)^2}$$

Now, since the square root is an increasing function, we may as well skip it, since it doesn't affect the choice of $\hat{\beta}$. The strategy will therefore be to choose $\hat{\beta}$ in such a way that the expression $||X_d\hat{\beta} - Y_{train}||^2$ is minimized. This expression is called the residual sum of squares (RSS). Note that we have the following:

$$RSS(\hat{\beta}) = ||X_d\hat{\beta} - Y_{train}||^2 = (X_d\hat{\beta} - Y_{train})^T(X_d\hat{\beta} - Y_{train})$$

$$=$$

$$\hat{\beta}^T X_d^T X_d\hat{\beta} - \hat{\beta}^T X_d^T Y_{train} - Y_{train}^T X_d\hat{\beta} + Y_{train}^T Y_{train}$$

$$=$$

$$\hat{\beta}^T X_d^T X_d\hat{\beta} - \hat{\beta}^T X_d^T Y_{train} - (\hat{\beta}^T X_d^T Y_{train})^T + Y_{train}^T Y_{train}$$

Note that $\hat{\beta}^T X_d^T Y_{train}$ is a scalar and so $\hat{\beta}^T X_d^T Y_{train} = (\hat{\beta}^T X_d^T Y_{train})^T$. We may therefore write

$$RSS(\hat{\beta}) = \hat{\beta}^T X_d^T X_d\hat{\beta} - 2\hat{\beta}^T X_d^T Y_{train} + Y_{train}^T Y_{train}$$

Note that this is a quadratic function of the vector $\hat{\beta}$. We now derive a closed-form solution for the $\hat{\beta}$ that minimizes the residual sum of squares in the special case where the design matrix $X_d$ has full

column rank, that is, when the columns of $X_d$ are linearly independent. This is called ordinary least squares regression (OLS) and the estimator for the coefficient vector is called the OLS-estimator.

**Theorem 1.1 (OLS-regression)**

If the design matrix $X_d$ is of full column rankt, then there is exactly one vector $\hat{\beta}$ that minimizes the residual sum of squares. $\hat{\beta}$ is given by

$$\hat{\beta} = (X_d^T X)^{-1} X^T Y_{train}$$

**Proof**

We consider $RSS(\hat{\beta}) = \hat{\beta}^T X_d^T X_d \hat{\beta} - 2\hat{\beta}^T X_d^T Y_{train} + Y_{train}^T Y_{train}$ as a differentiable function of $m+1$ variables $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m)$. The gradient $\nabla_{\hat{\beta}} RSS$ is given by

$$\nabla_{\hat{\beta}} RSS = 2X_d^T X_d \hat{\beta} - 2X_d^T Y_{train}$$

Setting the gradient to zero, we obtain

$$2X_d^T X_d \hat{\beta} - 2X_d^T Y_{train} = 0$$

$$\Leftrightarrow$$

$$X_d^T X_d \hat{\beta} = X_d^T Y_{train} \tag{2}$$

Now, since $X_d$ has full column rank, it must be the case that the $n \times n$ matrix $X_d^T X_d$ is invertible. We can therefore multiply both sides of equation (2)by $(X_d^T X_d)^{-1}$ to obtain

$$\hat{\beta} = (X_d^T X_d)^{-1} X_d^T Y_{train}$$

Hence $\hat{\beta} = (X_d^T X_d)^{-1} X_d^T Y_{train}$ is a critical point for $RSS$. Taking the second derivative $\nabla^2_{\hat{\beta}} RSS$, we obtain

$$\nabla^2_{\hat{\beta}} RSS = 2X_d^T X_d$$

Again, because of the full column rank condition on $X_d$, it must be the case that $X_d^T X_d$ is positive definite. It then follows that $\hat{\beta} = (X_d^T X_d)^{-1} X_d^T Y_{train}$ is a global minimum for $RSS(\hat{\beta})$. The proof is now complete. $\Diamond$

## 2.2 Robust standard errors

It is clear that the OLS-estimator $\hat{\beta} = (X_d^T X_d)^{-1} X_d^T Y_{train}$ is linear estimator. A linear estimator $\hat{\gamma}$ of some parameter $\gamma$ is said to be a best linear unbiased estimator, or BLUE (Wikipedia contributors 2025), if it satisfies the following conditions:

1. It is unbiased, that is, $\mathbb{E}[\hat{\gamma}] = \gamma$
2. It has the lowest sampling variance among the class of unbiased linear estimators

It can be shown that the OLS-estimator is BLUE if the following conditions are met:

1. The errors $\epsilon_i = y_i - X_{d,i}\hat{\beta}$ are uncorrelated, that is, $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$
2. $\mathbb{E}[\epsilon_i] = 0$
3. The errors are homoscedastic, that is, they all have the same finite variance, $Var(\epsilon_i) = \sigma^2 < \infty$ for all $i$.

When at least one of these conditions are not met, for instance, under heteroscedasticity, the OLS-estimator is no longer guaranteed to BLUE. As a consequence, since confidence intervals for the estimates of the coefficients rely on estimates of the variance of the errors, it may be the case that

1. The confidence intervals are too wide or too narrow.
2. The calculated p-values are incorrect

When there is reason to believe that there is heteroscedasticity, robust standard errors offer an alternative method to estimate the variance-covariance matrix of the OLS-estimator, which in turn gives trustworthy confidence intervals and p-values. Robust standard errors do not change the estimate of $\beta$ but rather adjusts the standard errors to give trustworthy confidence intervals and p-values.

## 2.3 Normality of errors

When the errors $\epsilon$ are normally distributed, it can be shown that the following holds:

1. The $\hat{\beta}$ estimates are normally distributed
2. the standardized statistic $t_i = \frac{\hat{\beta}_i - \beta_i}{\hat{SE}(\hat{\beta}_i)}$ follows a $t$-distribution with $n - m - 1$ degrees of freedom for all $n$.

This in turn enables us to calculate exact confidence intervals and p-values for any $n$. What happens when the errors are not normally distributed? Well, then the distributions of $\hat{\beta}$ and $t_i$ are no longer known and we have to rely on the central limit theorem to calculate approximate confidence intervals and p-values. In this setting, small sample inference is not theoretically justified. Strictly speaking, the inference is only valid in the limit, as $n$ tends to infinity. Normality of errors is therefore desirable but not crucial if the sample size is sufficiently large.

## 2.4 Linear relationship between the predictors and the target

When using linear regression, it is of course important that the relationship between the explanatory variables $X$ and the dependent variable Y are linear. Most often, this is not the case and then the model will obviously not generalize well. Moreover, for inference, it can bias the coefficients and invalidate confidence intervals and p-values leading to misleading conclusions.

To mitigate non-linearity, one can use polynomial regression, but another way to do it is to use splines (Wikipedia contributors 2025). Splines work by breaking up a continuous variable into piecewise polynomials that are joined smoothly at certain points called knots. So instead of using global polynomial terms, one uses different polynomials for different subsets of the values of the variable. This way, the result is a flexible curve that bends where it needs to bend, without overfitting.

## 2.5 Statistical tests and statistics

Besides the usual tests, plots and statistics involved in linear regression, such as the $R^2$, the $F$-statistic and VIF-values, a number of other statistical tests and statistics are mentioned in the method section below. Here we give a brief description of each of them.

### 2.5.1 The Pearson correlation coefficient

The Pearson correlation statistic $r$ (Wikipedia contributors 2025) estimates to what degree two numerical variables exhibit a linear relationship. Given samples $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$ of two variables $X$ and $Y$, It is defined as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

, where $\bar{x}$ and $\bar{y}$ are the averages of $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$ respectively. The statistic takes values in $[-1, 1]$. A value of -1 indicates a perfect decreasing linear relationship and a value of 1 indicates a perfect increasing linear relationship. A value of zero indicates no linear relationship.

### 2.5.2 The $\eta^2$ statistic

This statistic (Wikipedia contributors (2025)) estimates the proportion of variance in a dependent variable $Y$ that is explained by a categorical independent variable $X$. Given samples $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$, it is defined by:

$$\eta^2 = \frac{\sum_{j=1}^{k} n_j(\bar{y_j} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

, where

1. $k$ is the number of categories in the variable $X$
2. $n_j$ is the number of observations in category $j$
3. $\bar{y}$ is the overall mean
4. $y_j$ is the mean of category $j$

$\eta^2$ takes values in $[0, 1]$. A value of 0 indicates no relationship and a value of 1 indicates that all the variance in $Y$ is explained by $X$.

### 2.5.3 The Cramér's V statistic

The Cramér's $V$ statistic (Wikipedia contributors (2025)) estimates the strength of association between two categorical variables $X$ and $Y$. It is based on the $\chi^2$ test statistic (the $\chi^2$ test is a test for determining whether two categorical variables are independent). Given samples $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$, it is defined by

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

, where

1. $\chi^2$ is the test statistic in the $\chi^2$ test
2. $k = min(r, c)$ where $r$ is the number of rows and $c$ is the number of columns in the contingency table for the $\chi^2$ test.

### 2.5.4 The Durbin-Watson test

The Durbin-Watson test (Wikipedia contributors (2025)) is a statistical test for autocorrelation in the residuals of a linear model. Autocorrelation implies that the residuals are not independent. The

null hypothesis of this test is that there is no autocorrelation.

### 2.5.5 The Shapiro-Wilk test

The Shapiro-Wilk (Wikipedia contributors (2025)) test is a statistical test for whether sample data came from a normal distribution. The null hypothesis for this test is that the data came from a normal distribution.

### 2.5.6 The Breusch-Pagan test

The Breusch-Pagan test (Wikipedia contributors (2025)) is a statistical test for detecting heteroscedasticity in the residuals of a linear model. The null hypothesis for this test is that there is no heteroscedasticity.

# 3 Method

The following steps have been taken:

1. Data collection
2. Data cleaning
3. Splitting into train, validation and test data
4. Exploratory data analysis
5. Modeling for inference
6. Modeling for prediction

We will now account for each of these steps in more detail.

## 3.1 Data collection

We collected ads from the website Blocket (https://www.blocket.se). A group of about 12 people (including myself) collected about a hundred ads each and an excel file of observations was created. The datacollection workload was divided up by in which region of Sweden the car was offered for sale. The following restrictions were applied

1. The make of the car had to be Wolksvagen
2. The model year could be no earlier than the year 2000

All in all, the data collection went well and no major problems occurred.

## 3.2 Data cleaning

The data was read into R and the following cleaning measures were taken:

1. Spelling errors were corrected.
2. Capitalization was applied to categorical variables in order to ensure consistency.
3. Spaces and commas were removed from numerical variables.
4. Some of the numerical variables included units, which were removed.
5. Some observations had color values in the model column. These observations were removed.
6. Any observation with null values in a column was removed.

Before cleaning, there were 1204 observations in the dataset. After cleaning, there were 1193.

## 3.3 Splitting into train, validation, and test data

We shuffled the dataset and then used 70% of it for training, 15% for validation and 15% for test. This gives the following cardinalities for each of the datasets:

| Dataset | Cardinality | Percentage |
|---|---|---|
| Training | 835 | 70 |
| Validation | 178 | 15 |
| Test | 180 | 15 |

Figure 3.1: Cardinality of datasets

## 3.4 Exploratory data analysis

Both univariate and multivariate exploration was performed. We now account for each of these:

### 3.4.1 Univariate- and target interaction analysis

Here we describe the analysis made for the variables that were chosen for the final regression model. Extensive analysis and transformations were performed for the other variables as well but we won't go into it here.

#### Försäljningspris

This is the target variable for the regression. Plots of the distribution for this variable can be seen in Figure 3.4.4.1 below

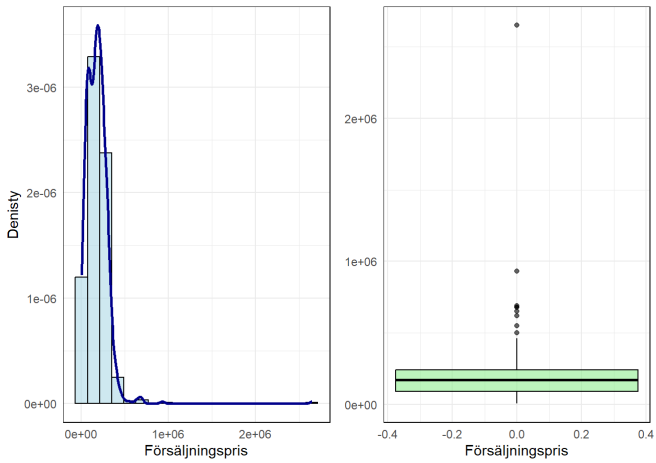


Figure 3.4.4.1: Distribution plots for Försäljningspris

As can be seen, there is skewness in the distribution and there is an extreme outlier. This outlier was removed and the variable was log-tranformed. The result of these actions can be seen in figure 3.4.4.2 below:

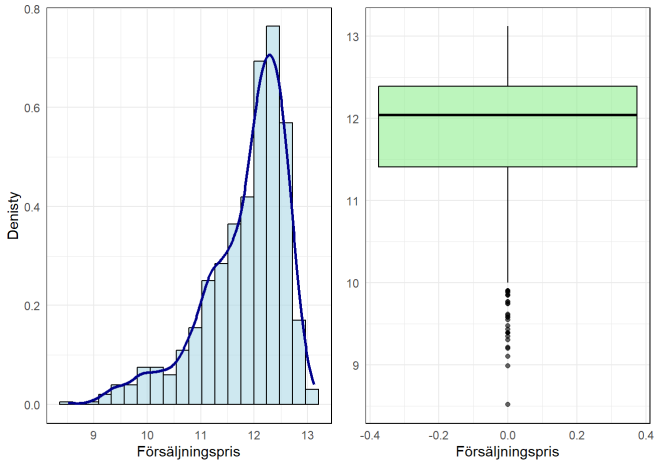


Figure 3.4.4.2: Distribution plots for Försäljningspris after transformations

**Miltal**

Plots of the distribution for this variable can be seen in Figure 3.4.4.3 below:

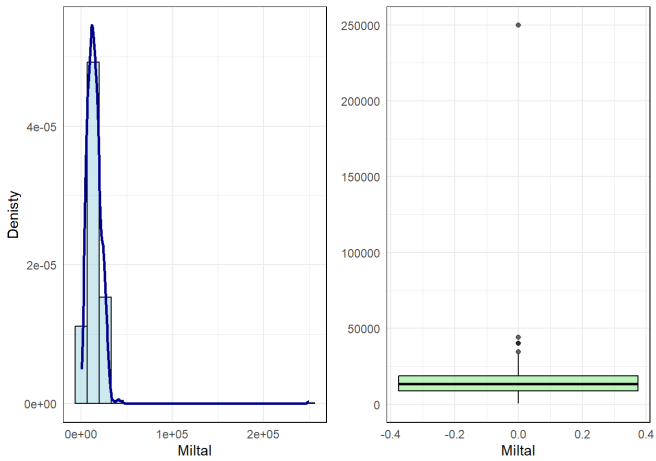


Figure 3.4.4.3: Distribution plots for Miltal

There is an extreme outlier, which was removed. The updated plot can be seen in Figure 3.4.4.4 below:

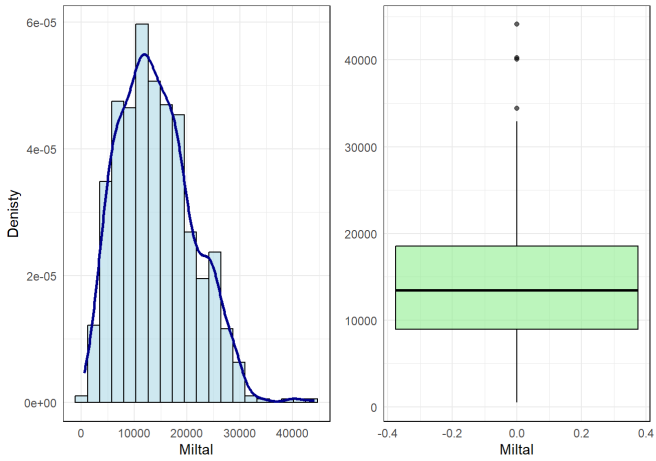


Figure 3.4.4.4: Distribution plots for Miltal after transformation

We also examined the interaction of Miltal and Försäljningspris (Figure 3.4.4.5 below):

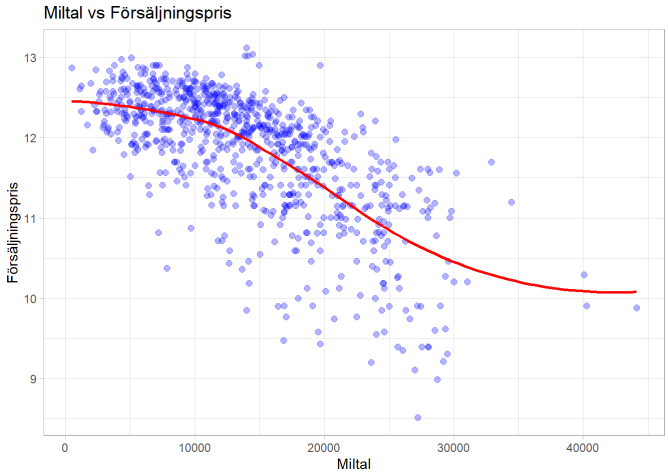


Figure 3.4.4.5: Interaction plot for Miltal and Försäljningspris

Here we see that the relationship is non-linear.

### Modellår

below are the plots for the variable Modellår:

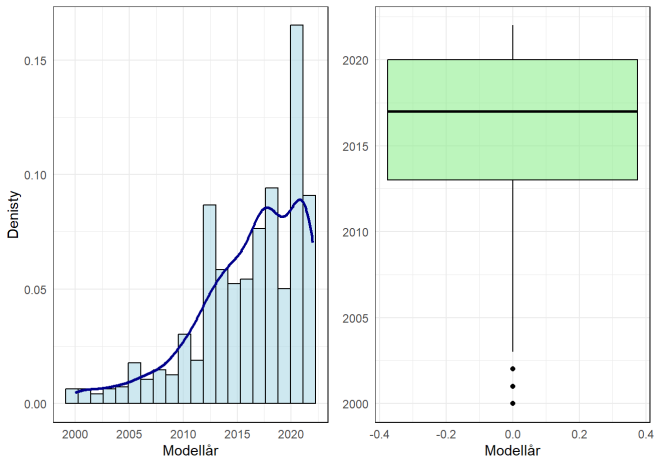


Figure 3.4.4.6: Distribution plots for Modellår

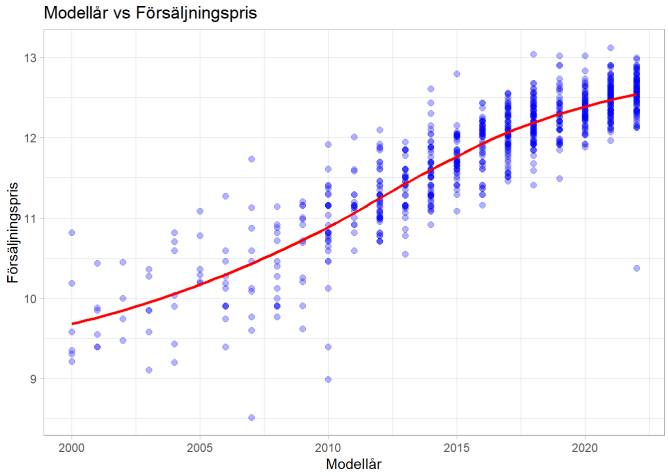


Figure 3.4.4.7: Interaction plot for Modellår and Försäljningspris

There are two things to notice here:

1. There seem to be only a few observations from before 2011. We checked that this was indeed the case (only one hundred observations for ten years worth of prices) and because of this we kept only observations where Modellår was no smaller than 2011
2. There is some non-linearity in the relationship between Modellår and Försäljningspris

### Hästkrafter (HK)

Below are the plots for the variable Hästkrafter (HK):

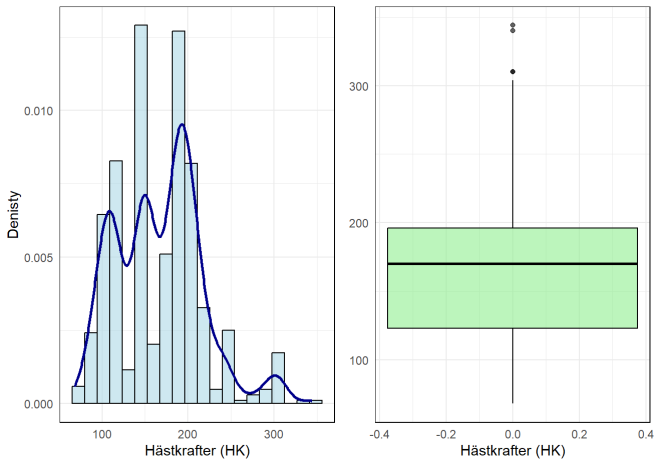


Figure 3.4.4.8: Distribution plots for Hästkrafter (HK)

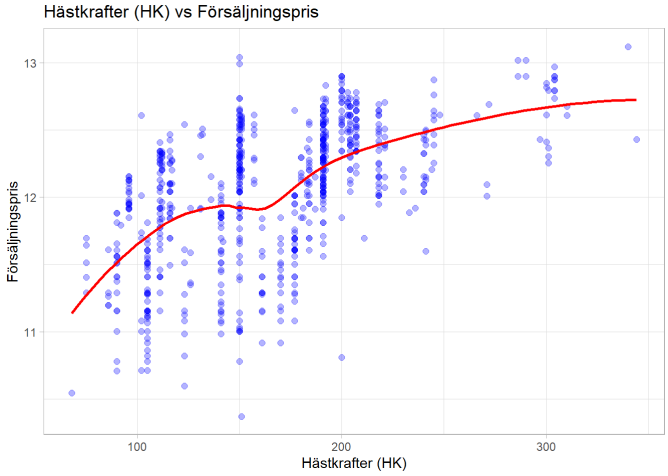


Figure 3.4.4.9: Interaction plot for Hästkrafter (HK) and Försäljningspris

The distribution plots looks acceptable but once again, the relationship between predictor and target is non-linear.

### Växellåda

Växellåda is a categorical variable (Manuell/Automat) but we can still plot the distribution and target interactions:

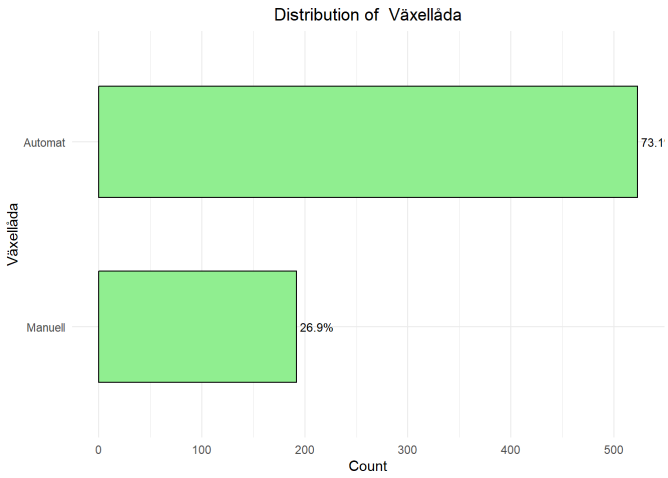


Figure 3.4.4.10: Distribution plot for Växellåda

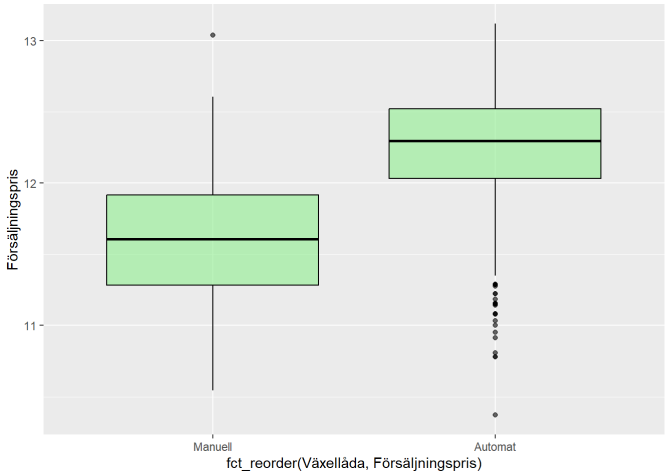


Figure 3.4.4.11: Interaction plot for Växellåda and Försäljningspris

This looks good. The imbalance isn't too severe and there is a clear difference in means of Försäljningspris for the two categories.

### 3.4.2 Multivariate analysis

For the multivariate analysis, we created an association matrix which can be seen in figure 3.4.2.1 below:

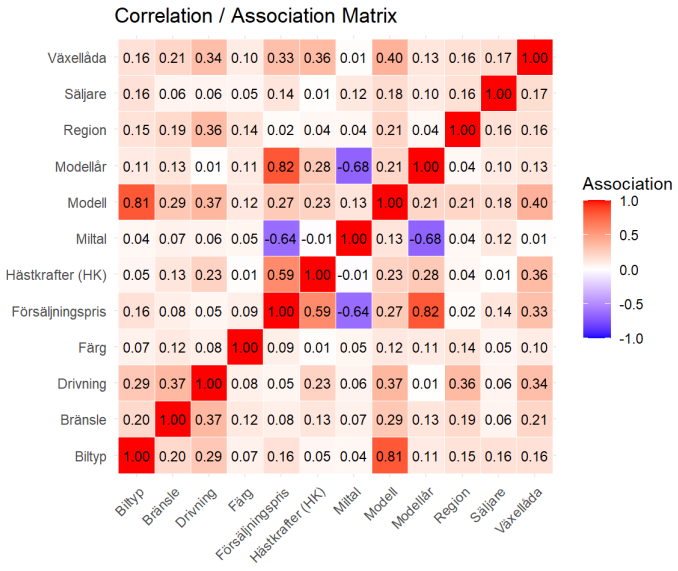


Figure 3.4.4.11: Association matrix for all variables

Here, the numbers are calculated as follows:

1. When comparing a numerical variable to a numerical variable, we use the ordinary Pearson correlation coefficient.
2. When comparing a numerical variable to a categorical variable, we use the $\eta^2$ statistic.
3. When comparing two categorical variables, we use the Cramér's V statistic.

From the matrix, we note the following:

1. The variables most associated with Försäljningspris are Modellår, Miltal, Hästkrafter (HK) and Växellåda
2. Miltal and Modellår are highly correlated

## 3.5 Modeling for inference

### 3.5.1 Feature selection

Following Occam's Razor, we started by selecting only features with an absolute association value larger than 0.3. This resulted in the following chosen variables:

1. Modellår
2. Miltal
3. Hästkrafter (HK)
4. Växellåda

With this decision, we expected to run into multicorrelation problems, especially since there is a strong correlation between Miltal and Modellår. It turns out that we had a stroke of luck and obtained very good results with this combination of features. So good, in fact, that we decided to not proceed any further with the feature selection.

### 3.5.2 Model fitting

We scaled the numerical variables and fitted a linear model using basis-splines of different degrees for some of the numerical variables. The best results occurred when using splines of degree 3 for Modellår and Hästkrafter (HK). No splines were needed for Miltal. We then proceeded to check VIF values. These can be seen in Figure 3.5.2.1 below:

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| bs(Modellår, df = 3) | 2.383898 | 3 | 1.155796 |
| bs(`Hästkrafter (HK)`, df = 3) | 2.109902 | 3 | 1.132514 |
| Växellåda | 1.986539 | 1 | 1.409446 |
| Miltal | 2.129874 | 1 | 1.459409 |

Figure 3.5.2.1: VIF values for the fitted model

We then proceeded to plot the histogram of residuals. The results can be seen in Figure 3.5.2.2 below:

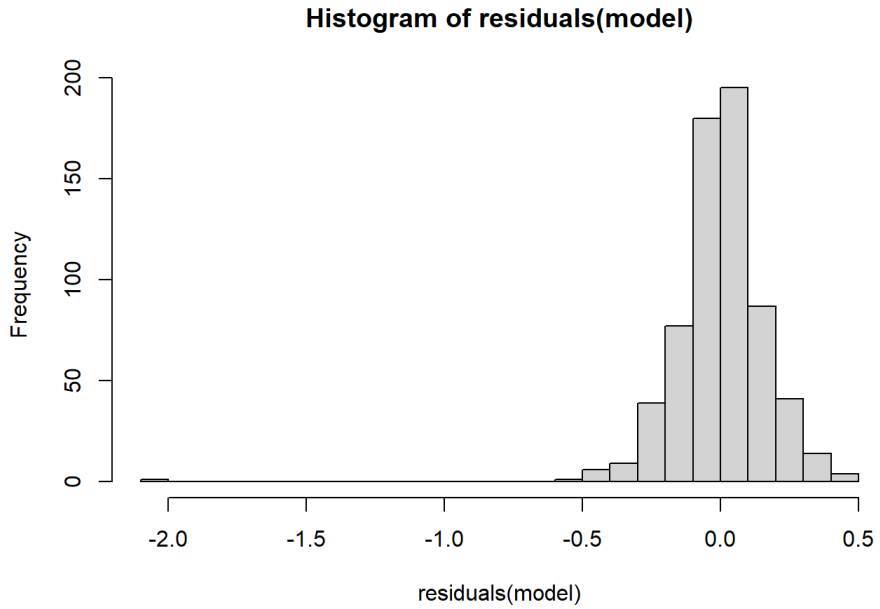


Figure 3.5.2.2: Histogram of residuals for the model

Here we see that there seems to be a single observation that kills the symmetry of the histogram. We located this observation (Figure 3.5.2.3):

A tibble: 1 × 12

| Försäljni... <dbl> | Säljare <fctr> | Bränsle <fctr> | Växellåda <fctr> | Miltal <dbl> | Modellår <dbl> | Biltyp <fctr> |
|---|---|---|---|---|---|---|
| 31900 | Företag | Diesel | Automat | 7860 | 2022 | SUV |

1 row | 1-7 of 12 columns

Figure 3.5.2.3: A problem observation

This is a SUV from 2022 with very low Miltal and extremely low price. This is likely a typo. We deem this observation to be a true outlier that needs to be removed from the training data. After removal and retraining, the histogram of residuals looks as follows (Figure 3.5.2.4):

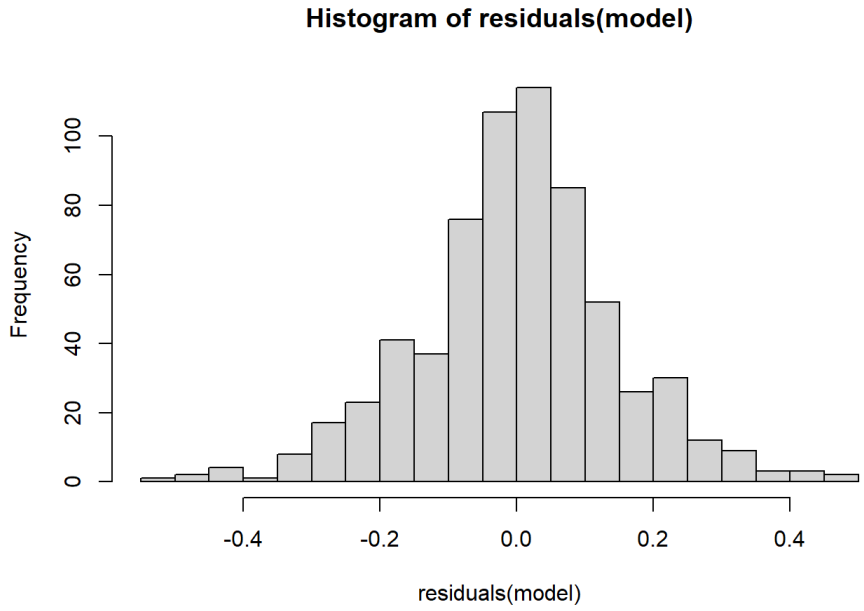


Figure 3.5.2.4: Histogram of residuals for the model after removal

This looks nice and symmetric, perhaps even normally distributed. The next step was to inspect the residuals vs fitted plot (Figure 3.5.2.5):

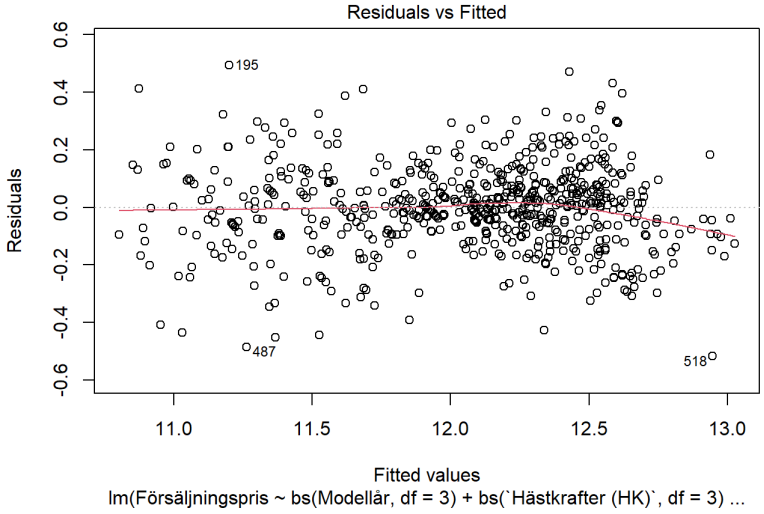


Figure 3.5.2.5: Residuals vs fitted plot

This plot looks like it is supposed to. A random cloud with no discernible patterns and a mean that looks close to zero. To confirm, we ran a Durbin-Watson test for autocorelation and calculated the mean residual. The results can be seen in figure 3.5.2.6 below:

```
        Durbin-Watson test

data:  model
DW = 2.0854, p-value = 0.8736
alternative hypothesis: true autocorrelation is greater than 0

Mean residual value: -4.01720226079361e-13
```

Figure 3.5.2.6: Output of Durbin-Watson test and mean residual calculation

This confirms what we saw in the plot. There is no autocorrelation and the mean residual is very close to zero. There were three flagged observations in the residuals vs fitted plot. We examined these and none of them warranted removal, so we kept them. Next, we inspected the Q-Q plot (Figure 3.5.2.7) to asses normality:

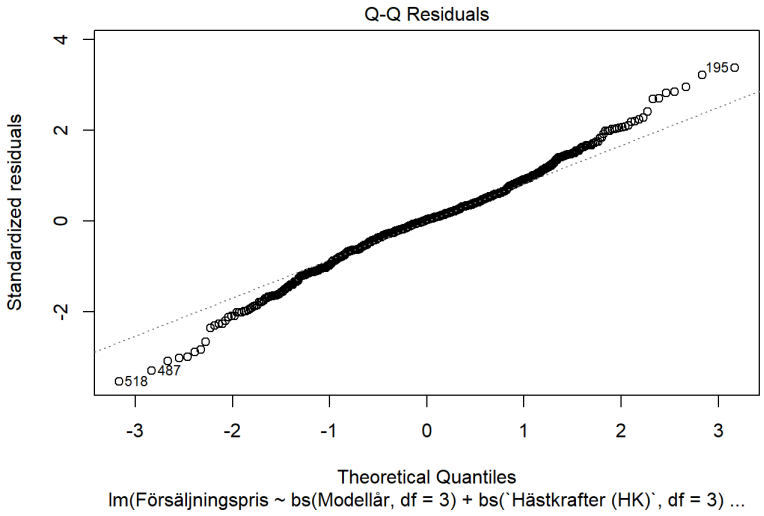


Figure 3.5.2.7: Q-Q plot

This plot suggests mild non-normality of the residuals. To confirm, we ran a Shapiro-Wilk normality test on the residuals. The result can be seen in Figure 3.5.2.8 below:

```
        Shapiro-Wilk normality test

data:  residuals(model)
W = 0.98997, p-value = 0.000196
```

Figure 3.5.2.8: Result of Shapiro-Wilk test

This confirms what we thought when looking at the Q-Q plot. The null hypothesis that the residuals are normally distributed is rejected. Since we have about 800 observations in the training data, the central limit theorem will work it's magic and this is likely not a problem for inference but we must acknowledge that any confidence intervals and p-values are non-exact. Next, we examined the Scale-Location plot (Figure 3.5.2.9):

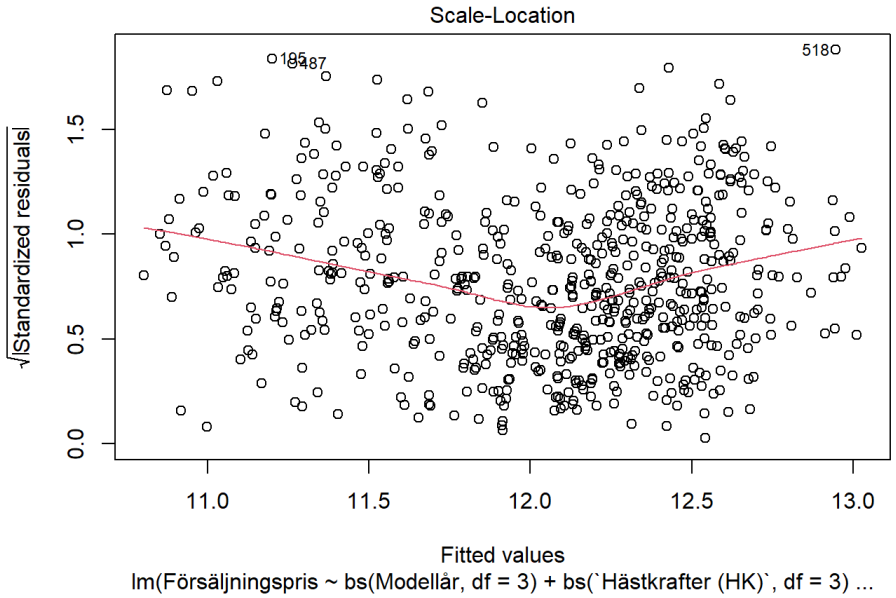


Figure 3.5.2.9: Scale-location plot

There is a slight U-shape present, indicating heteroscedasticity. To confirm, we ran a studentized Breusch-Pagan test, the output of which can be seen in Figure 3.5.2.10 below:

```
        studentized Breusch-Pagan test

data:  model
BP = 48.274, df = 8, p-value = 8.756e-08
```

Figure 3.5.2.10: Output of studentized Breusch-Pagan test

This confirms that there is indeed heteroscedasticity in the residuals. We will therefore use robust standard errors when we proceed to inferencing. Next, we inspected the Residuals vs leverage plot (figure 3.5.2.11):

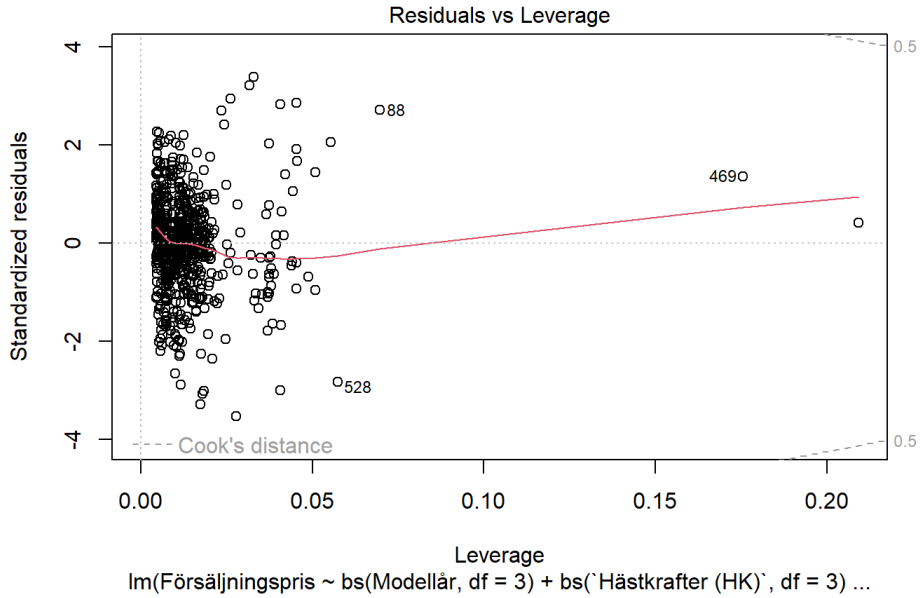


Figure 3.5.2.11: Residuals vs leverage plot

This looks reasonable. No major leverage points. The next step was to do the actual inferencing. To achieve this we calculate:

1. t-statistics and p-values for the coefficients using robust standard errors
2. the F-statistic using robust standard errors
3. confidence intervals for the coefficients using robust standard errors
4. The $R^2$ and adjusted $R^2$ for the model

The result is displayed in figure 3.5.2.12 below:

```
t test of coefficients:

                                Estimate Std. Error  t value  Pr(>|t|)
(Intercept)                   10.9469895  0.0650974 168.1633 < 2.2e-16 ***
bs(Modellår, df = 3)1          0.3983820  0.0948907   4.1983 3.067e-05 ***
bs(Modellår, df = 3)2          0.7868100  0.0438546  17.9413 < 2.2e-16 ***
bs(Modellår, df = 3)3          0.8066186  0.0526577  15.3181 < 2.2e-16 ***
bs(`Hästkrafter (HK)`, df = 3)1  0.9270175  0.1283776   7.2210 1.460e-12 ***
bs(`Hästkrafter (HK)`, df = 3)2  0.5744170  0.0770293   7.4571 2.868e-13 ***
bs(`Hästkrafter (HK)`, df = 3)3  1.2633866  0.1011803  12.4865 < 2.2e-16 ***
VäxellådaManuell              -0.1426279  0.0196717  -7.2504 1.195e-12 ***
Miltal                        -0.1759534  0.0098576 -17.8496 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Wald test

Model 1: Försäljningspris ~ bs(Modellår, df = 3) + bs(`Hästkrafter (HK)`,
    df = 3) + Växellåda + Miltal
Model 2: Försäljningspris ~ 1
  Res.Df Df      F    Pr(>F)
1    644
2    652 -8 634.98 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                                  2.5 %     97.5 %
(Intercept)                   10.8191608 11.0748182
bs(Modellår, df = 3)1          0.2120494  0.5847147
bs(Modellår, df = 3)2          0.7006947  0.8729253
bs(Modellår, df = 3)3          0.7032170  0.9100201
bs(`Hästkrafter (HK)`, df = 3)1  0.6749282  1.1791067
bs(`Hästkrafter (HK)`, df = 3)2  0.4231581  0.7256759
bs(`Hästkrafter (HK)`, df = 3)3  1.0647035  1.4620696
VäxellådaManuell              -0.1812564 -0.1039994
Miltal                        -0.1953102 -0.1565965
R-squared: 0.9134
Adjusted R-squared: 0.9123
```

Figure 3.5.2.12: Inferencing output

We also tested the splined variables as a whole for significance. The output of that test is shown in figure 3.5.2.13 below:

```
Linear hypothesis test:
bs(Modellår, df = 3)1 = 0
bs(Modellår, df = 3)2 = 0
bs(Modellår, df = 3)3 = 0

Model 1: restricted model
Model 2: Försäljningspris ~ bs(Modellår, df = 3) + bs(`Hästkraft
    df = 3) + Växellåda + Miltal

Note: Coefficient covariance matrix supplied.

  Res.Df Df     F    Pr(>F)
1    647
2    644  3 195.7 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Linear hypothesis test:
bs(`Hästkrafter (HK)`, df = 3)1 = 0
bs(`Hästkrafter (HK)`, df = 3)2 = 0
bs(`Hästkrafter (HK)`, df = 3)3 = 0

Model 1: restricted model
Model 2: Försäljningspris ~ bs(Modellår, df = 3) + bs(`Hästkraft
    df = 3) + Växellåda + Miltal

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    647
2    644  3 154.02 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.5.2.13: Significance tests for splined variables

Moreover, we created marginal effect plots for each of the predictors:

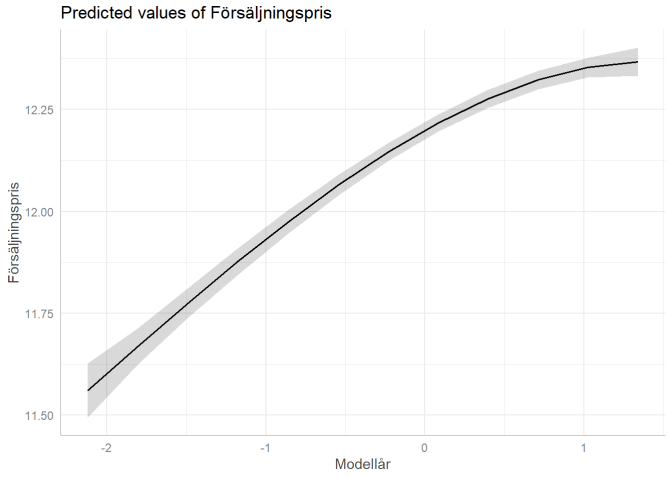


Figure 3.5.2.14: Marginal effect plot for Modellår

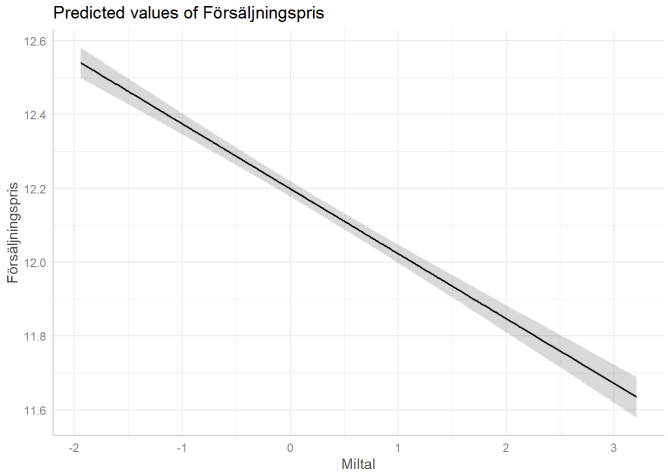


Figure 3.5.2.15: Marginal effect plot for Miltal

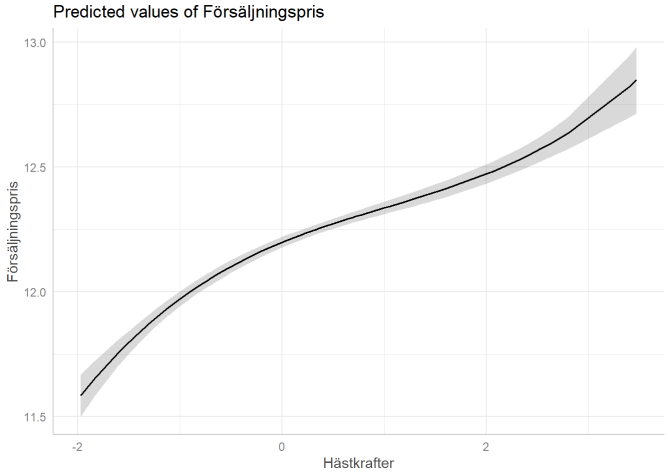


Figure 3.5.2.16: Marginal effect plot for Hästkrafter (HK)

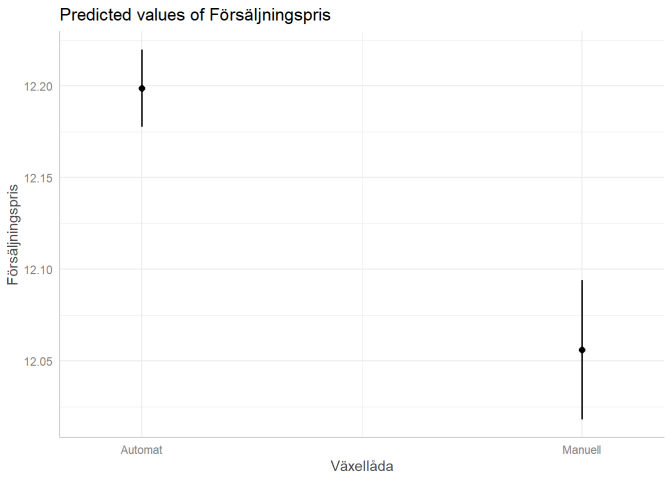


Figure 3.5.2.17: Marginal effect plot for Växellåda

Finally, we calculated the partial $R^2$ for each predictor to get a sense of the effect size of each variable (figure 3.5.2.18)

```
Modellår partial r2: 0.518836708962048
Hästkrafter (HK) partial r2: 0.501836653260499
Miltal partial r2: 0.40079407098405
Växellåda partial r2: 0.0837652226886687
```

Figure 3.5.2.18: Partial $R^2$for the predictors

## 3.6 Modelling for prediction

To examine the predictive power of our linear model, we predicted the validation data using our chosen model and a model without Miltal (since it had high correlation with Modellår). We calculated a "naive RMSE" for a model that simply predicts the mean of the target. We then calculated the RMSE and RMSE/mean_target for both of the models. The output of all these operations can be seen in figure 3.6.1 below.

| Mode | RMSE | RMSE/mean_target |
|---|---|---|
| Naive Model | 100669.96 | 0.6 |
| Chosen linear model | 33307.78 | 0.2 |
| Linear model without Miltal | 40644.03 | 0.24 |

Figure 3.6.1: Validation metrics for different models

Having chosen the final model (the linear model with Miltal included) we retrained it on the combined training- and validation data and then evaluated it on the test data to get an unbiased estimate of

the performance metrics. We got a RMSE of 38149.55 and a RMSE/mean_target of 0.2. We also calculated prediction intervals by bootstrapping the combined training- and validation data 1000 times and then evaluating on the test set. We then plotted the true response values against the mean of predictions and the prediction intervals (figure 3.6.2):

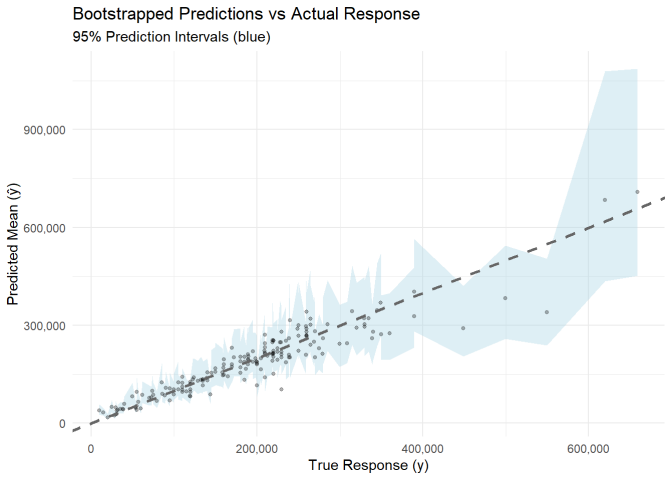


Figure 3.6.3: Prediction intervals

Next, we plotted the true responsse values values against the relative interval width (Interval width / mean prediction) (figure 3.6.4):

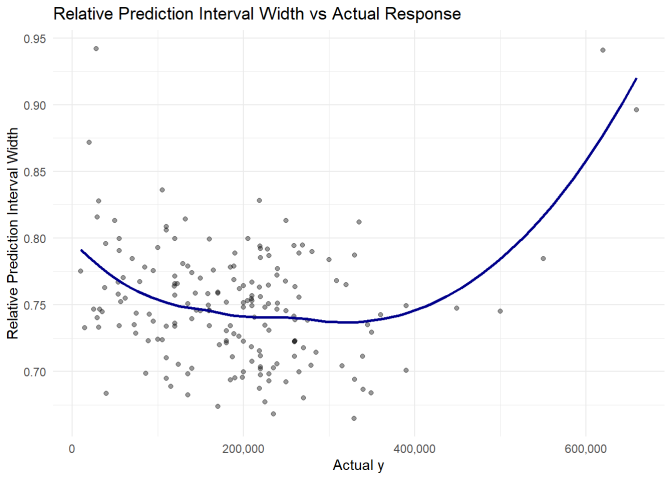


Figure 3.6.4: Relative width of prediction intervals

Finally, to compare the performance of a simple linear regression model to a more complex and

powerful model, we first trained an Xgboost regressor on the dataset with only the variables chosen for the linear model, and then with all the available variables. The metrics involved can be seen in figure 3.6.5 below:

| Mode | RMSE | RMSE/mean_target |
|---|---|---|
| Xgboost with restricted number of variables | 43651.12 | 0.23 |
| Xgboost with all variables | 36089.65 | 0.19 |

Figure 3.6.5: test metrics for different Xgboost models

# 4 Results

## 4.1 Modelling for inference

Figure 3.5.2.12 and 3.5.2.13 above show that the coefficients for all variables have the highest three-star significance whether we consider the splined variables separately or as a whole. As can be expected, and seen from inspecting figure 3.5.2.14 - 3.5.2.17 above, we can draw the following conclusions:

1. Modellår has an increasing effect on Försäljningspris
2. Miltal has a decreasing effect on Försäljningspris
3. Hästkrafter (HK) has an increasing effect on Försäljningspris
4. Växellåda Automat is associated with higher Försäljningspris than Växellåda Manuell

In figure 3.5.2.18 we see Modellår has the strongest effect on Försäljning, tightly followed by Hästkrafter (HK). Miltal has a smaller effect than Modellår and Hästkrafter (HK) but still a strong one. Växellåda has a more moderate effect on Försäljningspris.

## 4.2 Modelling for prediction

Figure 3.6.1 shows that even though Miltal and Modellår are highly correlated, Miltal does contribute to the performance metric for predictions, as can be seen in figure 3.6.1 above. It is clear from figure 3.6.2 and 3.6.3 that the model performs best when the true values of Försäljningspris are in the mid range. After values of the response variable of about 400 000, the width of the prediction intervals start to get quite high. This is due to two things:

1. There are very few observations (only 21) with Försäljningspris above 400 000
2. We removed observations above 500 000 to get a model better suited for inference.

Finally, as can be seen in figure 3.6.5 above, The highly capable Xgboost model actually performs slightly worse when trained on the same dataset as the chosen linear model. Moreover, when the Xgboost regressor is given access to all the variables, it does perform better but only by a small amount.

# 5 Analysis and discussion

At the beginning of this paper the following three research questions were asked:

1. Which sub collection, if any, of these variables significantly explain the variance in price for these ads?
2. Can we train a linear model that predicts unsean data at an $rmse/mean(price)$ of at most 0.2?
3. Seeing as a linear model is highly biased, do we see a dramatic improvement in $rmse/mean(price)$ if we instead train a highly complex model, such as an xgboost regressor on this data?

For the first question, here is indeed a subset of variable that significantly explain the variance of price. These are:

1. Modellår (an increasing effect)
2. Miltal (a decreasing effect)
3. Hästkrafter (HK) (an increasing effect)
4. Växellåda (automatic transmisson has higher prices)

It is not at all surprising that these variables would have an effect on the price of a car. It is, however, a bit surprising that so few variables account for so much of the variance in Försäljningspris. The lesson learned is that sometimes less is more and Occam's razor applies.

The answer to the second question is yes. We can indeed produce a linear model that meets the metric requirement. An RMSE/mean_target of 0.2 is not excellent but it's not catastrophically high either, especially since the number of predictors is small and the dataset size is on the lower side. The result reflects a solid balance between accuracy and interpretability.

The third research question has a negative answer. We do not see a dramatic improvement in performance when training an Xgboost model on this data. This shows that when the circumstances are right, a carefully crafted simple linear model can be on par with a complex black box model. A simple model with high interpretability is always preferable to a black box model if the performance metrics are the same.

The greatest takeaways from the work done are:

1. We don't always need many predictors to get descent results
2. We don't always need complex models to get descent results

# 6 References


1. Statistics Sweden (2025). Personbilar i trafik, antal efter region, ägarkategori, tabellinnehåll och år.
2. Statistics Sweden (2025). Indikatorer inkomstfördelning efter region, inkomstslag, tabellinnehåll och år.
3. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) An Introduction to Statistical Learning: With Applications in R. 2nd edn. New York: Springer.
4. Wikipedia contributors (2025). Gauss–Markov theorem
5. Wikipedia contributors (2025). Heteroskedasticity-consistent standard errors
6. Wikipedia contributors (2025). Spline (mathematics)
7. Wikipedia contributors (2025). Pearson correlation coefficient
8. Wikipedia contributors (2025). Effect size
9. Wikipedia contributors (2025). Cramér's V
10. Wikipedia contributors (2025). Durbin–Watson statistic
11. Wikipedia contributors (2025). Shapiro–Wilk test
12. Wikipedia contributors (2025). Breusch–Pagan test

# 7 Theoretical questions

### Question 1

Describe briefly what a Quantile-Quantile plot is.

### Answer

Consider a (increasingly sorted) sample $(x_1, ..., x_n)$ from some random variable $X$ with unknown distribution. For each $i \in \{1, ..., n\}$, let $q_i$ be the $\frac{i}{n+1}$-th quantile of the standard normal distribution. A QQ-plot is a scatterplot of the pairs $(q_i, x_i)$. If the sample came from a normal distribution, the scatterplot would be close to a straight line. Hence a QQ-plot is a way to visually check if a sample came from a normal distribution.

### Question 2

Your colleague Karin asks you the following: "*I've heard that in machine learning, the focus is on prediction, whereas in statistical regression, you can do both prediction and statistical inference. What is meant by that? Can you give any example?*". What is your answer to Karin?

### Answer

It is true that in machine learning, the focus is on obtaining a good model for prediction. But in statistical regression, one is not only interested in having a model that performs well on prediction tasks. One is also interested in which of the features contribute to the predicted value and by how much. Suppose, for instance, that we had a dataset where the variable we're predicting is wages and then we have some large number of explanatory variables for each observation. In machine learning we would train many models, perhaps using all the variables, and once we found a model that performs well on the test set, we would be satisfied. But in statistical regression, we want to discard any variable (perhaps first name if that was a predictor) that doesn't contribute to the variance of wages in any meaningful way. Moreover, for the variables that remain, we want to estimate the relative contribution between variables. Does level of education contribute more to the variance of wages than age, or is it the other way around? If the model assumptions are satisfied, one can answer such questions using statistical analysis. As a side note, there is a concept in machine learning called explainability which entails some of the elements of statistical regression/learning but it lacks the same rigor that is present in statistical learning. An example of where explainability is needed is if a bank has a machine learning model for predicting whether a customer will default on a loan. If the bank denies the loan, then it needs to be able to explain why the model made the prediction that it made.

**Question 3**

What is the difference between confidence intervals and prediction intervals for predicted values?

**Answer**

Consider a (multiple) linear model $y = \epsilon + \beta_0 + \sum_{i=1}^{n} \beta_i x_i$. A confidence interval for a prediction $\hat{y}$ only takes into account the uncertainty in the estimation of the coefficients $\beta_i$, whereas a prediction interval takes into account both the uncertainty in the estimation of the coefficients and the uncertainty in the noise variable $\epsilon$. Prediction intervals are therefore always at least as wide as, and often wider than confidence intervals.

**Question 4**

The multiple linear regression model may be written as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

How are the parameters interpreted?

**Answer**

Geometrically, the equation can be interpreted as a $p-1$ dimensional hyper plane in $p$-dimensional euclidean space. The $\beta_0$ coefficient is the $Y$-value where the hyper plane intersects the $Y$-axis. The other coefficients can be interpreted as the slopes of the hyper plane along the coordinate axes. For a non-geometrical interpretation, $\beta_0$ can be considered a base value and then the other $\beta_i$ can be interpreted as: "*If we keep all other variables $x_{j \neq i}$ fixed and increase $x_i$ by one unit, then $Y$ is increased by $\beta_i$*".

**Question 5**

Your colleague Nils asks you the following: "*Is it true that in statistical regression modelling, one doesn't have to use training, validation and test sets if one uses measures such as BIC? What's the logic behind this?*" What is your answer to Hassan?

**Answer**

At the risk of offending Nils by giving Hassan the answer instead of Nils, I would say the following: It is not a hundred percent true, but there is a grain of truth to it. For instance, in the limit, as the number of training observations tend to infinity, AIC produces the same value as leave-one-out cross validation, and BIC also exhibits some nice asymptotic properties. But there are alot of caveats and

assumptions involved and it is definitely sample size dependent. Therefore, since computing power is not a problem nowadays, train-val-test or cross validation is a more attractive alternative.

**Question 6**

Explain the algorithm for best subset selection.

**Answer**

The best subset selection algorithm is a brute force algorithm for selecting which predictors (if any) to include in a model. Given $p$ available predictors and some integer $k \in \{1, ..., p\}$ there are clearly $\binom{p}{k}$ distinct $k$-combinations of predictors. Each these combinations corresponds to one model that utilizes these predictors for training. For each $k$, there is at least one model, say $\mathcal{M}_k$, which has the largest $R^2$ among the $k$-combinations. In the set $\{\mathcal{M}_0, ..., \mathcal{M}_p\}$, where $\mathcal{M}_0$ is the naive model predicting the mean response, there is at least one model $\mathcal{M}_{best}$ with the best generalization ability, as measured by some appropriate metric or procedure. The best subset selection algorithm simply searches through the entire search space to find $\mathcal{M}_{best}$. For each $k \in \{1, ..., p\}$, $\binom{p}{k}$ models are considered. Including the naive model $\mathcal{M}_0$, we can use the binomial theorem to calculated the total number $m$ of models considered as follows:

$$m = 1 + \sum_{k=1}^{p} \binom{p}{k} = \sum_{k=0}^{p} \binom{p}{k} = 2^p$$

Hence the best subset selection algorithm has exponential time complexity, making it unfeasible except for when the number of predictors, $p$, is small.

**Question 7**

The following is a quote from the statistician George Box: "*All models are wrong, some are useful*". Explain the meaning of that quote.

**Answer**

The quote means that any statistical model is merely an approximation of the phenomena it aims to model. Some approximations are closer to reality than others and therefore they are useful. The linear regression model is an example of a statistical model which can be useful if the underlying relationship between predictors and response is approximately linear. The quote is not only true of statistical models, however. In physics, Newton's model of gravity as a force between objects with mass was useful enough to take us to the moon. But it was ultimately wrong, since it cannot account for the observations of the orbit of Mecury around the sun. For that we need the general theory

of relativity, which models gravity, not as a force but as geodetic paths on a Lorentzian space-time Manifold, the curvature of which, is determined by objects with mass and vice versa in the sense that space-time tells mass how to move and mass tells space-time how to bend. This model is even more useful than the Newtonian one, but ultimately, probably wrong too since it is incompatible with quantum mechanics and breaks down in singularities (black holes, big bang).

# 8 Self evaluation

**Question 1**

What has been the most fun in this assignment?

**Answer**

The most interesting part was handling the problems that occurred, such as non-normality and heteroscedasticity

**Question 2**

How have you handled challenges? What are the lessons learned?

**Answer**

I've handled challgenges by not giving up and by searching for the information that I need to solve the problems.

**Question 3**

Which grade do you think you deserve and why?

**Answer**

I think I deserve VG because I fullfil each of the criteria for that grade.

**Question 4**

Anything you want to bring up with Antonio?

**Answer**

No.