Statistical Analysis on BMW Car Dataset of Blocket

Submitted to: Antonio Prgomet

Prepared by: Garima Choudhary

Date: 26 March, Kunskapskontroll_R

Contents: POC and Gropparbete frågor

Problem Statemen

"Investigating the Factors Influencing the Selling Price of BMW Cars: A Statistical Analysis"

Here we are trying to explore and analyze the various factors(variables)like Milage, Model Year, that affect the selling price of BMW cars using statistical methods. For this we are going to create multi linear regression, and afterwords we are trying to select the best model. It highlights the need to identify and understand the key determinants of pricing in the second-hand BMW car market, which can aid in decision-making for buyers and sellers alike.

Approach

1. Data Collection: Utilizing web scraping techniques, I retrieved data from the "Blocket" website and stored it in a CSV file format. This enables seamless loading and reading of the data into R for further analysis and processing.

2. Library Loading: Necessary libraries such as `readxl`, `MASS`, `leaps`, `car`, `Metrics`, `broom`, and `ggplot2` are loaded into the environment to facilitate data handling, analysis, visualization, and model assessment.

3. Data Loading: The Excel file containing the BMW car dataset is read using the `read_excel()` function from the `readxl` package. The dataset is stored in the **`cars`** object for further processing.
4. Data Preprocessing: Any observations with missing values are removed from the dataset using the `na.omit()` function to ensure data completeness.

5. Linear Regression Modeling: Two linear regression models are fitted to the data: The first model (`logrr`) regresses selling price (`Pris`) on description (`Beskrivning`). The second model (`logr3`) regresses selling price on mileage (`Miltal`).

6. Model Selection: Backward, forward, and stepwise selection methods are employed to determine the most parsimonious models based on the Akaike Information Criterion (AIC).

```
# Backward selection
fulmdl <- lm(Pris ~ ., data = cars)
bkww <- step(fulmdl, direction = "backward", trace = 2)
```

```
# Forward selection
nulmd1 <- lm(Pris ~ 1, data = cars)
frww <- step(nulmd1, scope = list(lower = nulmd1, upper = fulmdl), direction = "forward", trace = 3)
```

```
# Stepwise selection
stepwise <- step(nulmd1, scope = list(lower = nulmd1, upper = fulmdl), direction = "both", trace = 3)
```

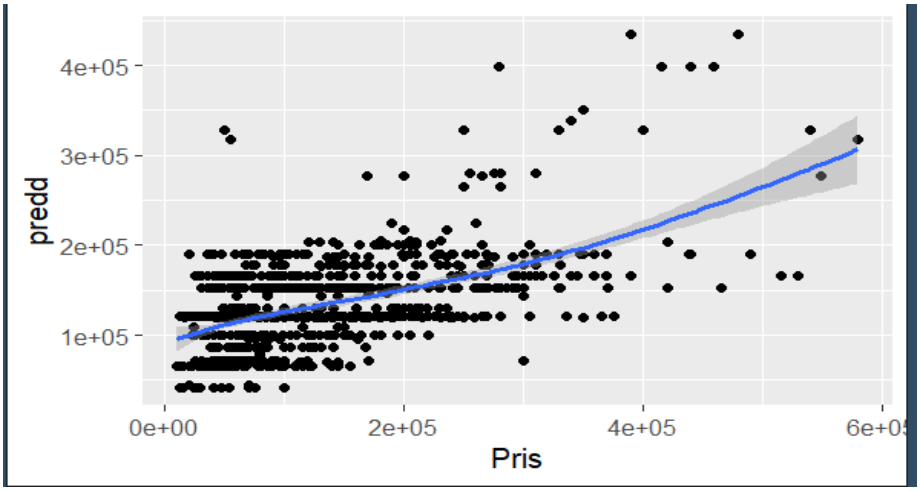
7. Plotting: Scatter plots and regression lines are generated to visualize the relationship between predicted and actual selling prices using both base R functions and `ggplot2`.

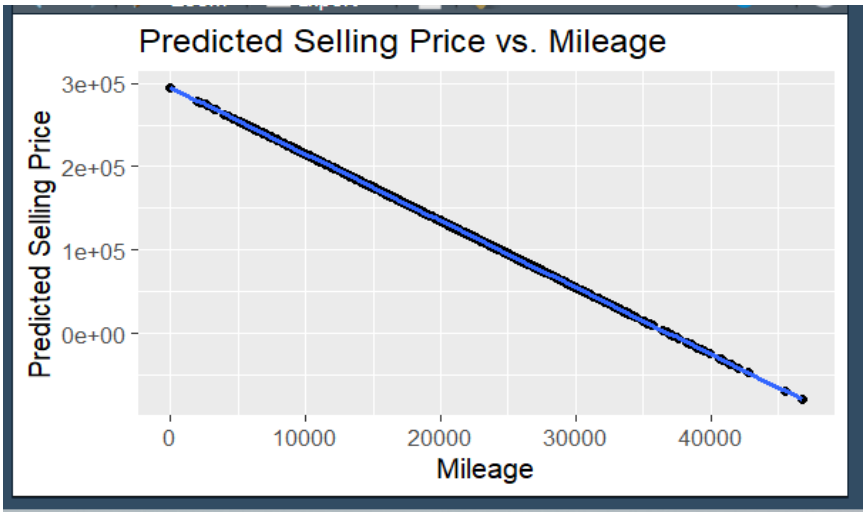
8. Broom Package: The `broom` package is utilized to conduct model assessment, including ANOVA and obtaining model summary statistics such as the R-squared value and residual standard error.

Data Source and Visualization Preview

Beskrivning	Län	Årsmodell	Drivmenel	Miltal	Växellåda	Modell	Pris	Link	Datum
BMW118	Västmanlan	2023	Bensin	2700.0	Automat	118	349000	https://www.blocket.se/annons/1400467158	27 feb. 14:57
BMW530	Skåne	2023	Hybri	2029.0	Automat	530	515000	https://www.blocket.se/annons/1400489467	22 feb. 19:34
BMW520	Hallan	2022	Diesel	7159.0	Automat	520	465000	https://www.blocket.se/annons/1400563351	16 mars 16:47
BMW330	Stockholm	2022	Hybri	3400.0	Automat	330	489000	https://www.blocket.se/annons/1400640649	I lörags 20:22
BMW530	Uppsala	2022	Hybri	2750.0	Automat	530	529000	https://www.blocket.se/annons/1400513556	20 mars 19:12
BMW520	Göteborg	2021	Diesel	17631.0	Automat	520	285000	https://www.blocket.se/annons/1400516801	6 mars 19:44
BMW330	Göteborg	2021	Hybri	12200.0	Automat	330	347000	https://www.blocket.se/annons/1400628682	3 apr. 08:25
BMW330	Jönköping	2021	Hybri	7306.0	Automat	330	350000	https://www.blocket.se/annons/1400536667	8 mars 19:28
BMW330	Gävleborg	2021	Hybri	7200.0	Automat	330	347500	https://www.blocket.se/annons/1400615015	30 mars 19:06
BMW745	Stockholm	2021	Hybri	7186.0	Automat	745	579900	https://www.blocket.se/annons/1400589777	24 mars 13:03
BMW520	Stockholm	2021	Diesel	7000.0	Automat	520	420000	https://www.blocket.se/annons/1400650225	I tisags 17:30
BMW330	Västmanlan	2021	Hybri	6878.0	Automat	330	310000	https://www.blocket.se/annons/1400600588	26 mars 20:46

Data frame





Hypothesis Testing

Null Hypothesis (H0): There is no relationship between mileage and selling price; the coefficient of Miltal (β1) is equal to zero.

Alternative Hypothesis (H1): There is a significant relationship between mileage and selling price; the coefficient of Miltal (β1) is not equal to zero.

The null hypothesis (H0) posits no relationship between mileage and selling price (β1=0), while the alternative hypothesis (H1) asserts a significant relationship (β1≠0). With a highly significant p-value (< 2.2e-16), the null hypothesis is rejected, confirming a substantial association between mileage and selling price.

The coefficient estimates of -8.008 suggests that, holding other variables constant, each unit increase in mileage results in an $8.008 decrease in selling price.

Conclusion

The analysis conducted on the relationship between mileage and selling price of BMW cars yielded significant results. The null hypothesis, which suggests no relationship between mileage and selling price ($\beta 1=0$), was strongly rejected, supported by a highly significant p-value (< 2.2e-16). This confirms a substantial association between mileage and selling price, aligning with the alternative hypothesis (H1) proposing a significant relationship ($\beta 1\neq 0$).

The coefficient estimates of -8.008 indicates that for every unit increase in mileage, the selling price decreases by $8.008, holding other variables constant. This finding underscores the importance of mileage as a predictor of selling price in the BMW car market.

Model diagnostics revealed that approximately 29.41% of the variance in selling price can be explained by mileage, as indicated by the multiple R-squared (0.2941) and adjusted R-squared (0.2683). The F-statistic (11.37) further supports the model's overall significance. Additionally, the residual standard error (79890) suggests the model's precision, while the AIC value (24903) indicates its relative fit compared to alternative models.

Furthermore, the model utilizing 'Beskrivning' as a predictor suggests a constant selling price ($134934) regardless of description. ANOVA results confirm a highly significant effect of 'Beskrivning' on 'Pris', emphasizing the predictive power of this variable in determining selling price.

Overall, these findings underscore the significance of mileage as a determinant of selling price in the BMW car market, providing valuable insights for buyers and sellers alike.

Grupparbete frågor

1.Vem har du arbetat i grupp med?

Svar: Jag arbetade med Ahamad, Mustafa, Cristopher, Anna och Lina i grupp.

2. Hur har ni i gruppen arbetat tillsammans?

Svar: Vi diskuterade om webbskrapning, olika variabler och modellval.

3.Vad var bra i grupparbetet och vad kan utvecklas?

Svar: Det var ett tillfälle att arbeta med faktiska datamängder med andra som faktiskt arbetssenario och jag tycker att vi borde diskutera oftare med i gruppen.

5.Vilka är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?

Svar: det ger dig en bild av det faktiska arbetsscenariot.

6.Finns det något du skulle ha gjort annorlunda? Vad i sådana fall?

Svar: Jag har inte upplevt något sådant faktiskt.