A Study of the Price of a Used Ford Car Using Possible Contributing Factors

Eric Yong

Dataset Discussion

- General Research Question
  - Can you use contributing factors to determine the price of a used Ford car?
- Ford used car dataset
- Dependent Variable - price of the car (y)
- Independent Variables
  - Age (2024 - year) (X1)
  - Mileage (X2)
  - Road tax (X3)
  - MPG (X4)
  - Engine Size (X5)
  - Transmission - Manual, Automatic (X6), Semi-Auto (X7)
  - Fuel type - Petrol, Diesel (X8), Other (X9)

Model Selection and Diagnostics

- Model necessities
- Significance tests
- Best subset algorithms
- Heteroskedasticity
- Normality
- Outliers / Points of Influence
- Multicollinearity

Model Remedies

- Box-cox procedure
- Weighted least squares
- Ridge regression
- Robust regression

Hypothesis Tests

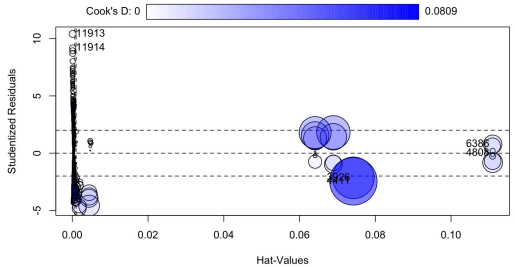
- F statistic
- T-value
- Confidence Interval

The Interaction Between Fuel Type and Road Tax on Price

Is the interaction between road tax (X_3) and engine type (Diesel - X_8, Other - X_9) on the price (Y) statistically significant?

Model Selection

- Price (Y)
- Road tax (X_3)
- Fuel type: Diesel (X_8), Other(X_9) (other: electric, hybrid, etc). Baseline: Petrol.
- Full model: $Y=\beta 3X3+\beta 8X8+\beta 9X9+\beta 38X3X8+\beta 39X3X9+\varepsilon$
- Reduced model: $Y=\beta 3X3+\beta 8X8+\beta 9X9$



Model Diagnostics

- Heteroscedasticity
  - Breusch-Pagan test: p-value < 2.2e-16
- Residual Normality
  - Shapiro-Wilk test: p-value < 2.2e^-16
- Outliers / Points of Influence
  - Influence plot (Cook's distance, studentized deleted residual, Hat value)
  - Indicates many points of influence

Model Remedies

Box-Cox

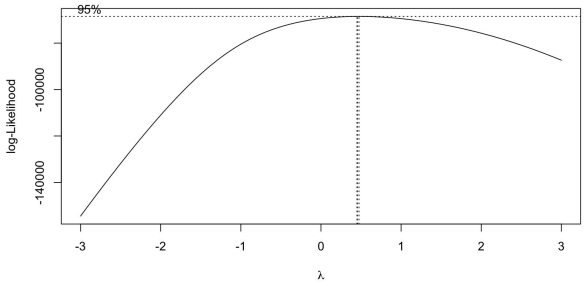
- Lambda: 0.4545
- Breusch-Pagan: p-value < 2.2e^-16
- Shapiro-Wilk: p-value < 2.2e^-16
- No improvement in R-square value
- Didn't work (made diagnostics worse)

WLS

- Breusch-Pagan: p-value = 1
- Shapiro-Wilk: p-value < 2.2e^-16
- R-squared of 0.3568 (improvement from 0.2498)
- Improved model!

Robust Regression

- Outliers and influence points are still present
- Cannot justify use of model.



$$\text{Price} = 8204.9329 + 28.5857*\text{tax} + 592.5001*\text{diesel} + 4670.2320*\text{other} + 19.4737*\text{tax}*\text{diesel} + 68.2862*\text{tax}*\text{other}$$

Final Model & Hypothesis Test

- Hypothesis Test
  - $H_0: \beta_{38}=\beta_{39}=0$ Ha: $\beta_{38}\neq 0$ or $\beta_{39}\neq 0$
  - ANOVA results in a F statistic of 304.35 which resulted in a p-value close to 0.
- Conclusion:
  - The interaction effect between fuel type and road tax on the Ford price is statistically significant.
  - 10-fold cross validation shows RMSE of 4106.679 and R-square of 0.2508439

The Effect of Transmission Type and MPG on Price

How does MPG (X_4) and transmission type (Automatic - X_6, Semi-Auto - X_7) jointly impact the price of used Ford cars in the UK? Is this impact significant?

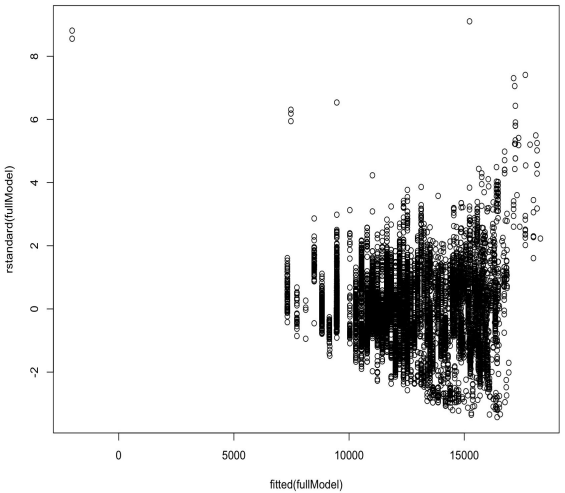
Model Selection

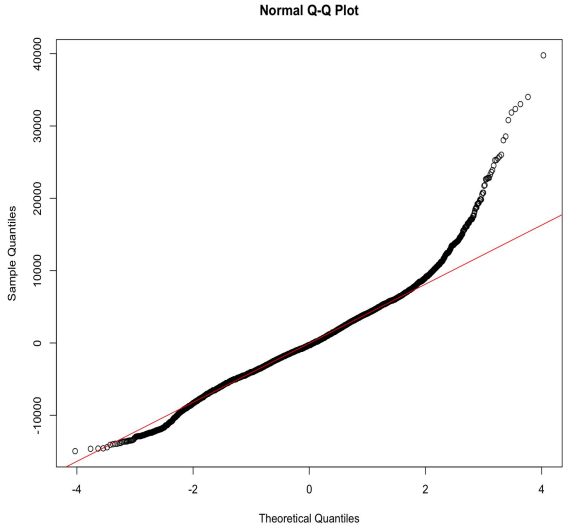
- Price (Y)
- MPG (X_4)
- Transmission type: Automatic (X_6), Semi-Auto(X_7)
- Best subset: adjusted R^2, Cp, AICp, and SBCp chooses model without Semi-Auto
- Stepwise regression: Also chooses model without Semi-Auto
- Full model: $Y=\beta 4X4+\beta 6X6+\beta 7X7+\beta 46X4X6+\beta 47X4X7+\varepsilon$

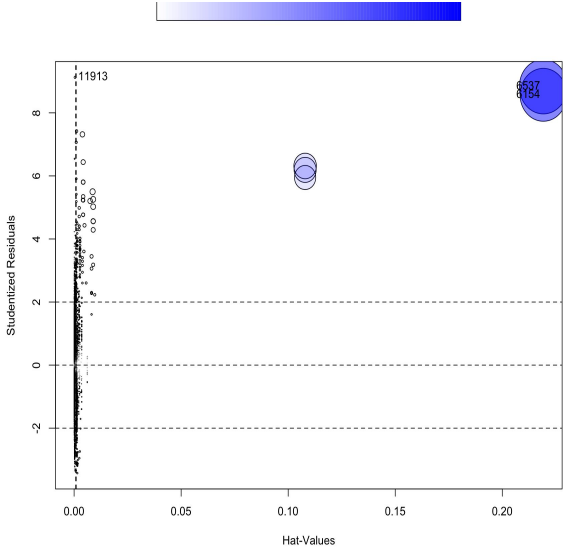
Model Diagnostics

- Heteroskedasticity
  - Residual plot
  - Breusch-Pagan test: p-value < 2.2e-16
- Normality
  - Normalized Q-Q plot
  - Shapiro-Wilk test: p-value < 2.2e^-16
- Outliers / Points of Influence
  - Influence plot

Plots


rstandard(fullModel)
fitted(fullModel)
8
6
4
2
0
-2
0
5000
10000
15000


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
40000
30000
20000
10000
0
-10000
-4
-2
0
2
4


11913
Studentized Residuals
Hat-Values
8
6
4
2
0
-2
0.00
0.05
0.10
0.15
0.20

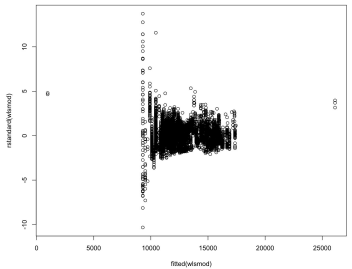
Model Remedies

Box-Cox

- Lambda: 0.5
- Breusch-Pagan: p-value < 2.2e^-16
- Shapiro-Wilk: p-value < 2.2e^-16
- Didn't work

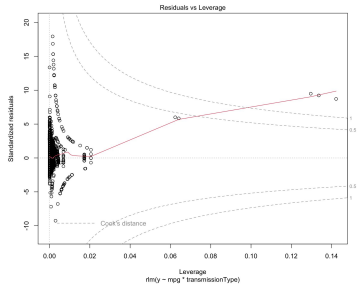
WLS

- Breusch-Pagan: p-value = 1
- Shapiro-Wilk: p-value < 2.2e^-16



Robust Regression

- Breusch-Pagan: p-value = 1
- Shapiro-Wilk: p-value < 2.2e^-16



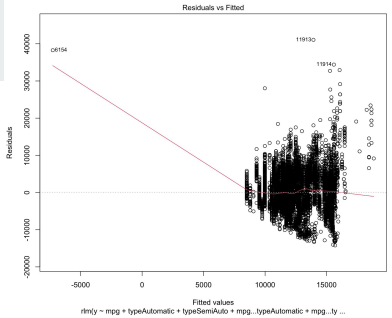
Final Model & Hypothesis Test

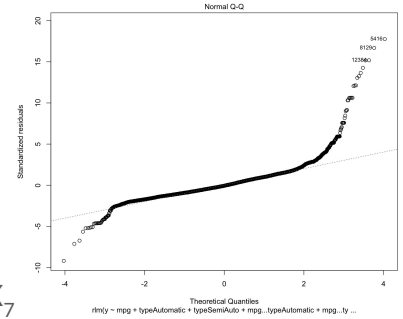
- Removed points of major influence after robust regression
  - Breusch-Pagan: p-value = 1
  - Shapiro-Wilk: p-value < 2.2e-16
- Hypothesis Test
  - H_0: $\beta_6=\beta_7=\beta_{46}=\beta_{47}=0$ Ha: $\beta_6\neq0$ or $\beta_7\neq0$ or $\beta_{46}\neq0$ or $\beta_{46}\neq0$
  - Full: $Y=17788.8998-105.0861X_4-1077.2596X_6+4048.1617X_7+77.3608X_4X_6-39.1338X_4X_7$
  - Reduced: $Y=X_0+\beta_6X_6$
  - Critical t-value: t(1-0.05/2, 17961-6) = 1.960096

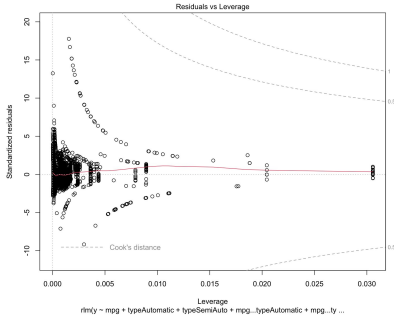
```
Call: rlm(formula = y ~ mpg + typeAutomatic + typeSemiAuto + mpg...typeAutomatic +
    mpg...typeSemiAuto, data = fixedFordData, weights = wts2)
Residuals:
      Min        1Q    Median        3Q       Max
-11.80118  -0.84995  -0.07957   0.88912  22.82313

Coefficients:
                    Value      Std. Error t value
(Intercept)       17788.8998   120.5163   147.6058
mpg                -105.0861     1.5983   -65.7506
typeAutomatic     -1077.2596   713.3484    -1.5101
typeSemiAuto       4048.1617  1018.1830     3.9759
mpg...typeAutomatic   77.3608    13.0781     5.9153
mpg...typeSemiAuto   -39.1338    18.5148    -2.1136

Residual standard error: 1.285 on 17955 degrees of freedom
```

```
t value
 147.6058
 -65.7506
 -1.5101
  3.9759
  5.9153
 -2.1136
```

Conclusion: Since all but one variable is significant, we reject the null hypothesis. Thus, the addition of transmission type does improve the model for predicting the price.







The Effect of Age and Mileage on Price

Is the effect of age (X1) and mileage (X2) on price (Y) statistically significant?

Model Selection

- Price (Y): Dependent variable representing the transformed vehicle price.
- Mileage and Age (X): Independent variables included to predict vehicle price.
- Transformation Applied: Box-Cox transformation with lambda = 0.4545 to address heteroscedasticity and non-normality issues.

Model Diagnostics

- Heteroscedasticity:
  - Assessed via the Breusch-Pagan test, indicating significant heteroscedasticity in the initial model.
- Residual Normality:
  - Evaluated using the Shapiro-Wilk test, revealing non-normal distribution of residuals in the initial model.

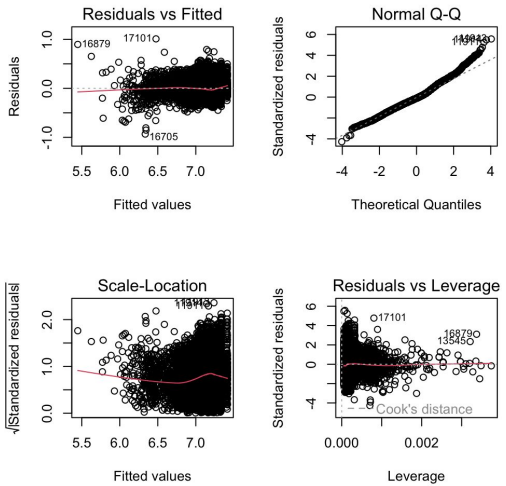
Model Remedies

Box Cox:

- Lambda: 0.5
- Breusch-Pagan Test:
- p-value < 2.2e^-16
- Shapiro-Wilk Test:
- p-value < 2.2e^-16
- Outcome:
- Transformation did not resolve heteroscedasticity or normality issues.

WLS:

- Breusch-Pagan Test:
- p-value = 1 (indicates the absence of heteroscedasticity)
- Shapiro-Wilk Test:
- p-value < 2.2e^-16 (residuals not normally distributed)
- Diagnostic Plot:
- (Insert diagnostic plot image) showing residuals post-WLS regression.



Robust Regression:

- Breusch-Pagan Test:
- p-value = 1 (model passes test for constant variance)
- Shapiro-Wilk Test:
- p-value < 2.2e^-16 (residuals not normally distributed)

Final Model and Hypotheses Test

Model equation:

$$Y = \beta_0 + \beta_{mileage} X_{mileage} + \beta_{age} X_{age} + \epsilon$$

Final Model:

$$Price_{transformed} = 7.728 - 1.685 * 10^{-6} * mileage - 0.07847 * age$$

Hypotheses testing:

$$H_0 = \beta_{mileage} = 0$$
$$H_0 = \beta_{mileage} \neq 0$$

$$H_0 = \beta_{age} = 0$$
$$H_0 = \beta_{age} \neq 0$$

```
> # Summary of WLS model
> summary(wls_model)

Call:
lm(formula = price_transformed ~ mileage + age, data = ford_clean,
    weights = weights)

Weighted Residuals:
      Min        1Q    Median        3Q       Max
-0.233417 -0.035271 -0.005627  0.031583  0.304440

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.728e+00  4.847e-03 1594.39   <2e-16 ***
mileage     -1.685e-06  8.645e-08  -19.50   <2e-16 ***
age         -7.847e-02  8.681e-04  -90.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05466 on 17961 degrees of freedom
Multiple R-squared:  0.5539,    Adjusted R-squared:  0.5538
F-statistic: 1.115e+04 on 2 and 17961 DF,  p-value: < 2.2e-16
```

The t-values for 'mileage' and 'age' are -19.50 and -90.39, respectively. We reject the null hypothesis.
The multiple R-squared value of 0.5539 indicates that approximately 55.39% of the variability in the transformed price can be explained by the model. Adjusted R-Squared is also similar at 0.5538. An F-statistic of 1.115e+04 and a p-value < 2.2e-16 indicates that the overall model is statistically significant.