DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KU LEUVEN

Report on project about regression

Statistical Data Analysis
Project 3

Mitja Mandić

May 2022

1 Introduction

For this assignment we switch from dataset on canteloupe melons to analysing the CO_2 emissions in grams per kilometer of cars from 2000 to 2013. As in the previous projects we select a random subsample of the data we do our study on, this time of size 500.

2 Full model analysis

Firstly, we draw a correlation matrix to check for potential multicollinearity in the data. Note that we also had to remove categorical variables `euro_standard`, `fuel_type` and `transmission_type` from the data to do perform the calculations.

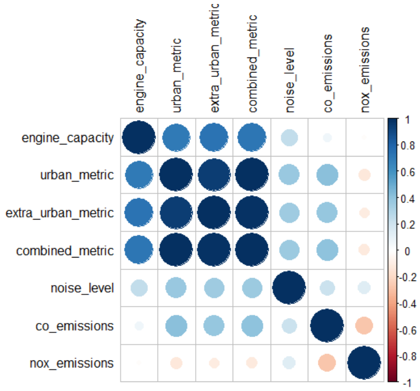


Figure 1: Correlation matrix of numeric predictors

In figure 1 we see that there is in fact some significant correlation between "metric" variables and engine capacity. Since combined metric is a weighted average of urban and extra-urban metric correlation between them is expected. Intuition also tells us that the larger engines consume more fuel, so corrleation to this variable is also not surprising. Calculating the correlation with the response variable, previously mentioned covariates are the most correlated with it. Determinant of the correlation matrix is almost zero, another proof that multicollinearity is present in our data. Below all values are presented.

```
co2  engine_capacity urban_metric extra_urban_metric combined_metric
1.00 0.75            0.98         0.97               0.99
noise_level co_emissions nox_emissions
0.40        0.34         0.02
```

We move on to fit the full model, meaning we include all the data as our covariates. While most of the coefficients in the model are statistically significant, `euro_standard4`, `urban_metric`, `extra_urban_metric`, `noise_level` and `co_emissions` all have p-values above 0.05.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.414e+00  5.252e+00   0.650   0.5160
euro_standard4           2.361e-03  4.482e-01   0.005   0.9958
```

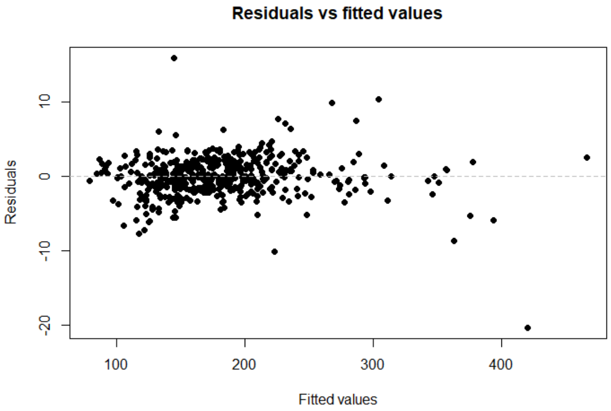
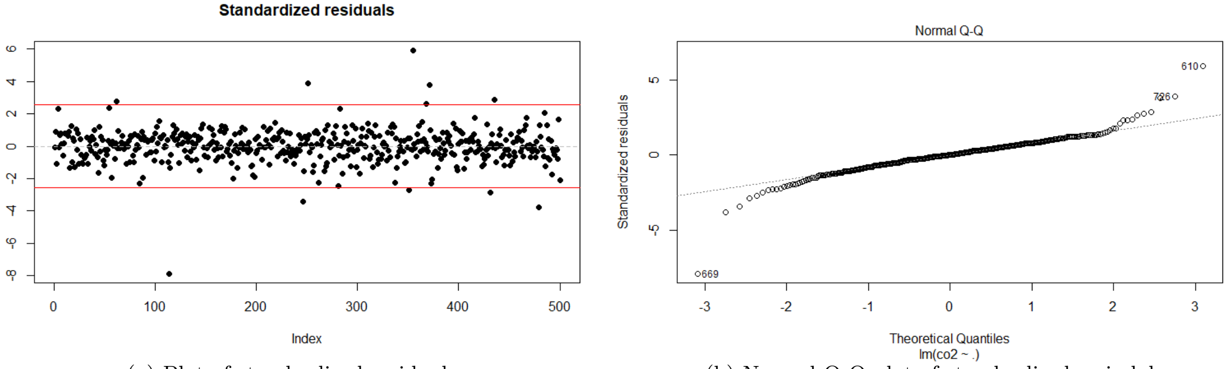


Figure 2: Residuals versus fitted value of the full model

```
euro_standard5          -3.469e+00  5.120e-01  -6.776 3.58e-11 ***
euro_standard6          -2.253e+00  1.001e+00  -2.250   0.0249 *
transmission_typeManual -1.199e+00  2.981e-01  -4.020 6.73e-05 ***
engine_capacity         -6.180e-04  2.962e-04  -2.087   0.0374 *
fuel_typeHybrid         -1.581e+01  7.172e-01 -22.041  < 2e-16 ***
fuel_typePetrol         -1.673e+01  6.301e-01 -26.557  < 2e-16 ***
urban_metric             5.363e-01  8.908e-01   0.602   0.5475
extra_urban_metric       2.673e+00  1.504e+00   1.778   0.0761 .
combined_metric          2.173e+01  2.386e+00   9.105  < 2e-16 ***
noise_level              1.281e-01  7.590e-02   1.687   0.0922 .
co_emissions            -3.631e-04  7.678e-04  -0.473   0.6364
nox_emissions            8.991e-03  2.244e-03   4.007 7.12e-05 ***
```

Even though a few variables are non-significant, the linear model assumptions are satisfied. The mean of errors is zero as is their sum, and their variance does not differ across the data. From the plot of errors versus fitted values we see that in the largest datacloud no trends are appearing. Fewer cars have very large fitted values of their CO_2 emissions and their residuals very a bit more. Some outliers are present, but in majority they are not severe.



(a) Plot of standardized residuals

(b) Normal Q-Q plot of standardised resiudals

Figure 3: Investigating behaviour of residuals of the full model

From the Q-Q plot in figure 3b we can conclude that the residuals normally distributed, as the majority fall on the diagonal line with some deviation towards the tails, with the three most notable outliers being Maserati Spyder (669 – very fancy sports car), Daewoo cars Matiz (610 – small family car from South Korea) and Volkswagen Touareg (label 726). This way we also confirm that Gauss-Markov conditions hold.

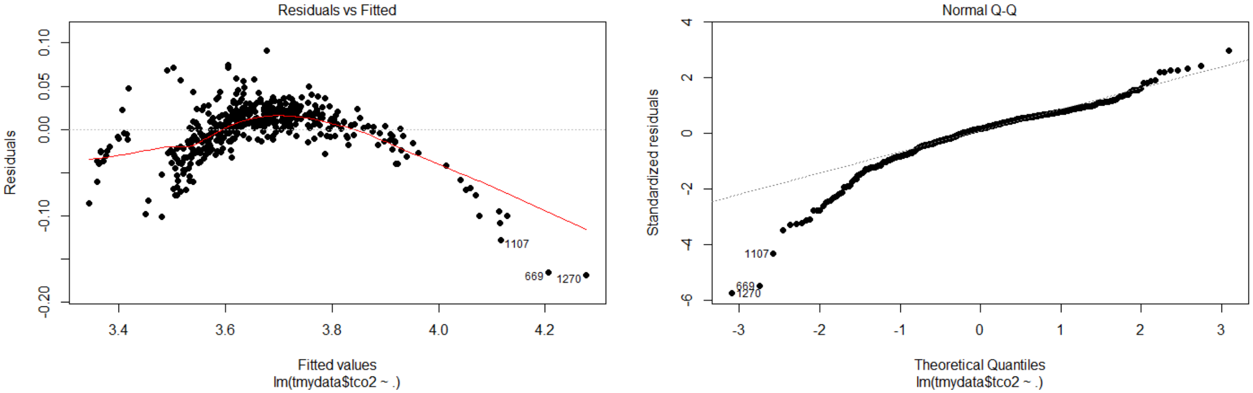
3 Transforming the response variable

In the following sections we turn our attention to potentially improving our model by transforming the response variable.

3.1 Box-Cox transformation

We start off by applying the Box-Cox transformation, for which we obtain $\lambda = -0.14$. This however does not yield satisfactoty results. The adjusted R^2 value remains high at 0.951, however the residuals do not exhibit trends of normal distribution. Several more outliers appear in the negative direction. The Q-Q plot shows heavier tails than in the model without a transformation and plotting resiudals versus fitted values shows a quadratic trend.

Box-Cox transformed response



(a) Plot of standardized residuals for the Box-Cox transformed response

(b) Normal Q-Q plot of standardised resiudals for the Box-Cox transformed variable

Figure 4: Investigating behaviour of residuals of the full model with Box-Cox transformed response variable

3.2 Logarithm transformation

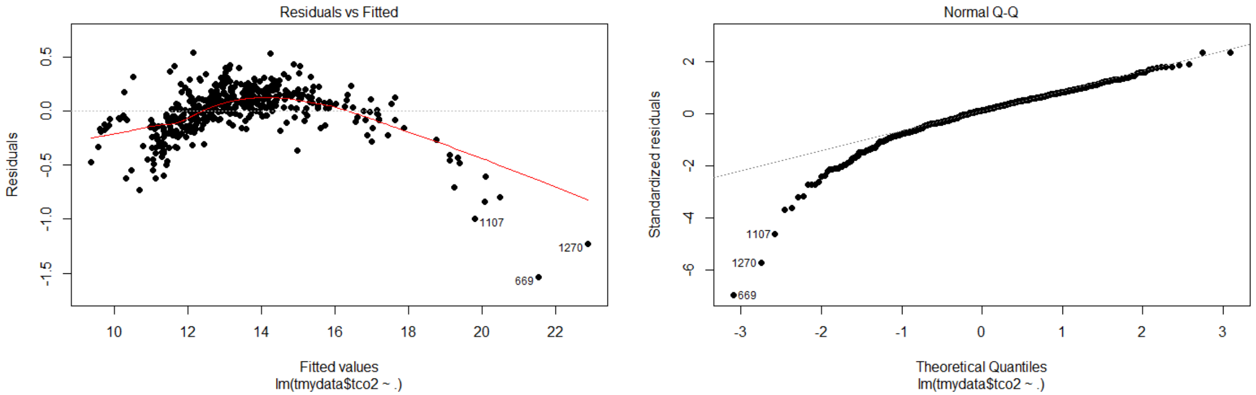
We proceed with a logarithm of the response variable. The conclusions are fairly similar to those of section , which is not surprising as the logarithm is simply a Box-Cox with parameter zero. In this particular case, 0 is even in the 95% confidence interval for λ in the previous section. As the plots look alike to those in figure 4 we do not include them here. We do not use the logarithm to transform the response either.

3.3 Square root

The last option we check is the square root. Once again we find that the transformation does not improve the fit. Residuals versus fitted value plot in figure 5a exhibits a quadratic trend, and especially error's the lower tail is much heavier than in the normal distribution, as seen in 5b.

The MSE of all models with a transformed response is lower compared to the original model, while also their adjusted R^2 remaining relatively high (lowest with Box-Cox transformation at 0.95). However, since their residuals behave in a strange manner any further inference with these models would be invalid. Therefore we conclude, that no transformation is needed for our data.

Square root of the response



(a) Plot of standardized residuals for the square root of the response

(b) Normal Q-Q plot of standardised resiudals for the square root of the response

Figure 5: Investigating behaviour of residuals of the full model with square root of the response variable

4 Variable selection

As mentioned in section 2 some variables in the full model appear to be statistically insignificant. We will now try to improve our model by eliminating them.

First we remove all variables with too high p-value, that is `urban_metric, extra_urban_metric, noise_level` and `co_emissions`, simultaneously (`euro_standard4` is part of a categorical variable). In this model all coefficients except for those corresponding to `euro_standard4` and `euro_standard6` are significant. Applying the F-test to check whether this can be done however rejects this hypothesis with a p-value of 0.0013.

This changes after we add `extra_urban_metric` to the new model. That way we do not reject the null that coefficients are zero and all covariates (except `euro_standard4`) are significant. The adjusted R^2 value remains very high at 0.9975 (compared to 0.9976 in the original model). Residuals behave appropriately and in fact rather similarly to the full model. Since omitted variables do not seem to add anything to the model we continue our project with the sparser one, now conisting of covariates: euro standard, transmission type, engine capacity, fuel type, extra urban metric, combined metric and NO_x emissions.

4.1 Can we remove euro_standard from the model?

Using the anova function on the reduced model, which still contains the variable euro_standard, we see in the results of the partial F-test that we cannot remove it from our model - the p-value obtained is essentially zero. Looking into the results in greater detail we see a very large p-value for euro_standard4, which implies that the slope for these cars is statistically the same as the baseline, which is euro_standard3. This is not the case for cars with engines of higher standards, as both of those have p-values below the 0.05. Therefore we cannot remove all slopes corresponding to euro_standard from our model simultaneously. In figure 6 we see that behaviour of standards 3 and 4 is very similar, while remaining different compared to engines with higher standards

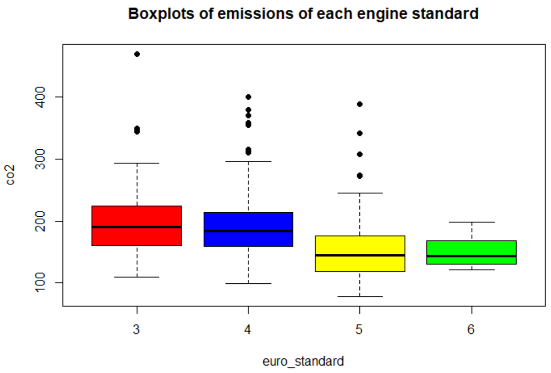


Figure 6: Similar trends visible between standards 3 and 4, and 5 and 6. Number of cars with engine standard 6 is however much smaller than the rest.

5 Confidence and prediction intervals

5.1 Confidence interval for β_1

Following the mentioned model, β_1 corresponds to the coefficient in front of euro_standard4. The 95% confidence interval is $\text{CI}(\beta_1, 0.95) = [-0.698, 0.961]$. We see that 0 is a possible option for this value, however removing all slopes related to euro_standard is still not possible as confidence intervals for other values do not include 0.

5.2 Prediction interval

Here we construct a 99% prediction interval for a car with the following specifications: euro_standard = 4, transmission_type = "Manual", engine_capacity = 2196, fuel_type = "Petrol", urban_metric = 9.2, extra_urban_metric = 5.6, combined_metric = 6.9, noise_level = 72, co_emissions = 273.5, nox_emissions = 43.

Prediction interval for its CO_2 emissions in grams per kilometer is $[156.35, 170.48]$, with the fitted value 163.42.