Statistical Data Analysis
Project 2

Jef Masereel, r0631651, group 19

June 22, 2020

1. Data inspection & preparation

1.1 Data inspection

Before starting with any regression analysis, it is important to start by properly inspecting the available data. The R script included with this report contains a detailed overview of the data contents through figures and data summaries, but to stay within the page limit most of these results are only referred to when deemed relevant for the topics of discussion in the following sections.

The given dataset consists of three different types of variables: categorical, standardized and positive continuous. The categorical predictor variables require separation into multiple binary variables, which will become relevant in the section on variable selection. The standardized predictor variables will be handled as normal proportion variables. Only the response variables are positive continuous (user counts). These differences should be kept in mind during selection of the appropriate methods, as will become clear in the next sections.

First a word on the univariate distributions: the normality assumption of predictor variables only seems (close to) valid for the variables Windspeed and Humidity. All other are skewed and/or have a distribution density with multiple peaks. This is confirmed by their p-values when applying the Shapiro-Wilkes test. The bivariate distributions were checked with a pairwise plot and inspection of the correlation values. The pairwise plot suggests light dependencies between most variables, but the one relation standing out is the high correlation between Temperature and Feelingtemperature. This linear relationship is confirmed by the correlation matrix. A more general approach to this is to check for multicollinearity in the given set of predictor variables. The included R script looks at VIF values and eigenvalues. The mean and max VIF values drop from (140,41) to (1.8,1.3) by simply leaving out the Temperature variable. This indicates the impact of the relation between the two temperature variables on the multicollinearity of this regression problem. The eigenvalues confirm this conclusion, their values can be reconstructed with the R script. That being said, the remaining variables still contain a small amount of multicollinearity, which might lead to computational issues for the regression coefficients. A well-considered variable selection will help to omit this problem, but first we'll consider which variables can be transformed to normality for more accurate regression.

1.2 Variable transformations

Starting with the predictor variables, it would be best if they all conformed to a normal distribution (with exception of the categorical variables of course). Given that these are standardized, some appropriate transformations to consider are logit, probit and asin(sqrt(P)) (as described in textbook, p93). The R script is used to compare the improvements by these different methods, and returns the table below. Based on these results, keeping previous findings into account, the choice was made to apply logit to both temperature variables and asin(sqrt(P)) to Windspeed. Humidity was left unchanged, since no meaningful improvements could be obtained.

variable	original	logit	probit	asin(sqrt)
Temperature	1.359506e-07	2.257233e-04	5.577747e-05	5.905972e-06
Feelingtemperature	5.460503e-06	9.058396e-04	4.466652e-04	1.047975e-04
Humidity	9.418048e-03	9.768850e-12	1.361361e-08	1.552829e-04
Windspeed	9.796789e-04	2.383762e-04	6.697307e-02	9.496965e-01

Table 1: Comparison of improvements obtained by different transformations

Concerning the response variables, trying a Box Cox transformation seems like the best option. For Casual, the normality assumption can be made valid by applying Box Cox with a lambda of 0.26. For Registered however, no good lambda value could be found so no power transform was applied to this predictor. After this, both predictors were standardized to zero-mean and a standard deviation of one. Besides generalizing the regression, this also makes it easier to compare performance results in later sections. Please note that these same values have to be applied to the test set when computing predictions with the final regression models.

1.3 Outlier detection

Now that these variables have undergone the appropriate transformations, it makes sense to inspect the dataset for outliers before starting the regression analysis. The R script uses robust regression on the full set of predictors and on the set of standardized predictors. This second regression allows us to generate full diagnostic plots, which a good visual overview of the different types of outliers present. As can be seen in the figures below, both response variables (separated datasets since the aim is to find one regression model for each) have a small amount of good and bad leverage points, as well as vertical outliers. The R script can be used for a more detailed inspection, but in conclusion 12 outliers were removed for Casual and 4 for Registered. This leaves our datasets with 318 and 326 observations respectively.

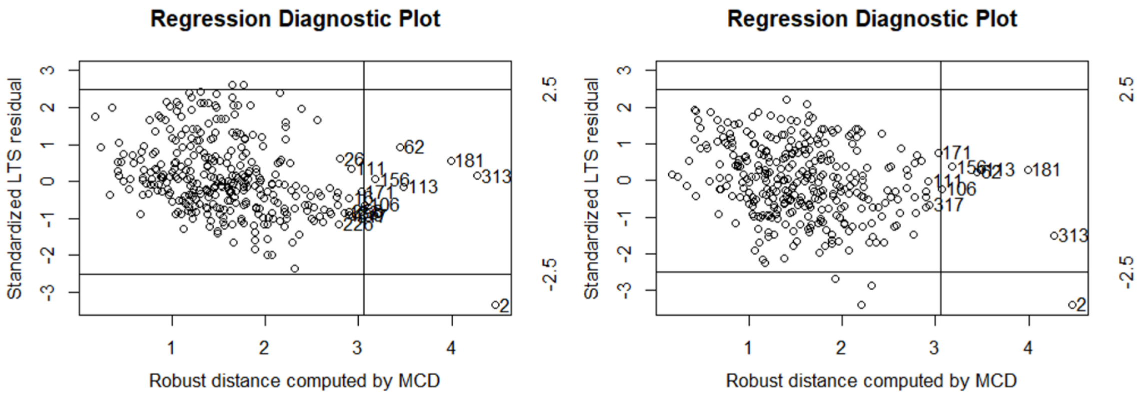

Figure 1: Robust outlier detection for casual (left) and registered (right)

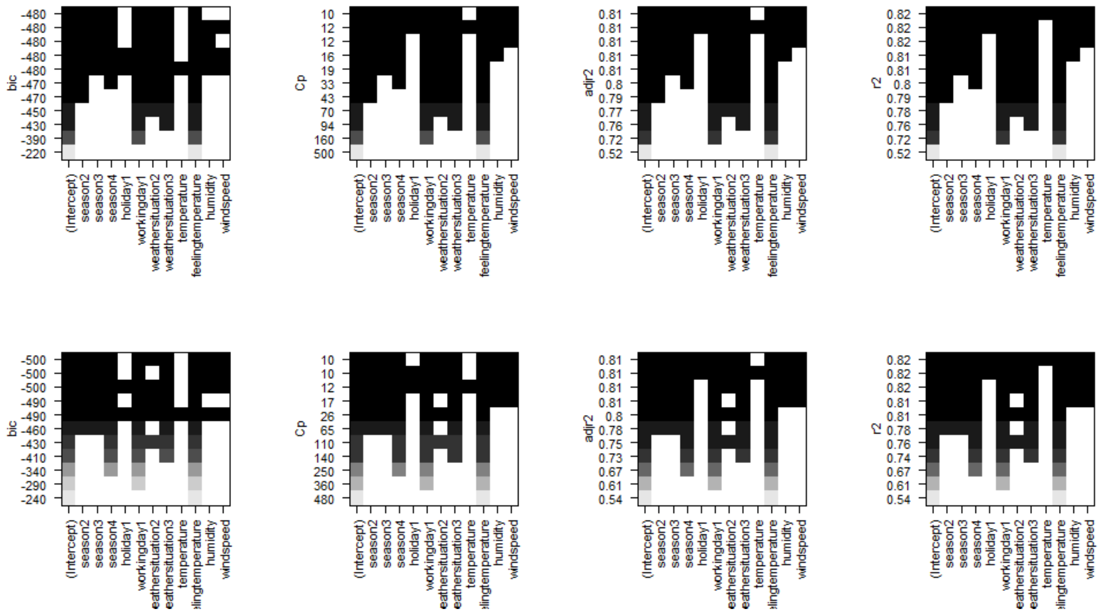

Figure 2: plot.regsubsets for Casual (top row) and Registered (bottom row)

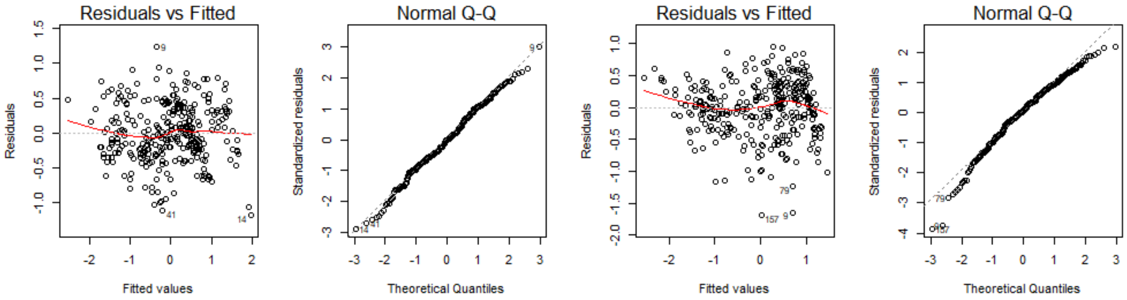

Figure 3: Residual plots of regression model for Casual (1st 2) & Registered (last 2)

2. Regression analysis

2.1 Variable selection

Based on the previous conclusions, the most important goals of this step are to reduce the issue of multicollinearity and select variables that are reasonably close to normality (if continuous). The R script goes well into detail on different metrics to find the most appropriate set of predictors for this regression problem. From figure 2 below it becomes apparent that they all follow a similar pattern. Since we only have a small number of observations, it is acceptable to go for a conservative selection. With all this in mind, the R script continues with the same selection of 4 predictor variables for both Casual and Registered: Season, Workingday, Weathersituation and Feelingtemperature. An interesting note is that seems to prefer the categorical variables over the standardized ones, even though in the given context it might seem like humidity and windspeed would be more accurate indicators of the weather in general.

2.2 Final model

Now that these decisions are made, the remaining regression analysis becomes quite straightforward. The R script shows the full implementation, but we'll shortly discuss the results of these regression models. Most important is to check whether the Gauss-Markov conditions are reached. Figure 3 shows residual plots for both models (Casual and Registered). It seems safe to assume that the residuals for Casual are approximately zero-mean and homoskedastic. The corresponding QQ plot also confirms normality. The R script returns a significant p-value of 0.37. For Registered however, the normality assumption is not as easy to confirm. Since no appropriate transformation was found for this response variable earlier on, some heteroskedasticity trasnferred into the model. Despite a low p-value, the QQ plot still look somewhat reasonable apart. It is to be expected that the second model will return a lesser performance on the test set.

2.3 Performance analysis

Before applying the model the test set must be prepped in exactly the same way as when building the regression model. The R script shows this implementation. Remember that the exact same mean and standard deviation values from the training set are used to standardize the response variables. No outliers were removed. The eventual performance metrics indicate are very bad for Registered, likely due to a mismatch in the distribution of obervations (strong difference in means and standard deviations). For Casual however, the results are relatively accurate given the low insight in contextual mechanisms. Please see the R script for detailed results.

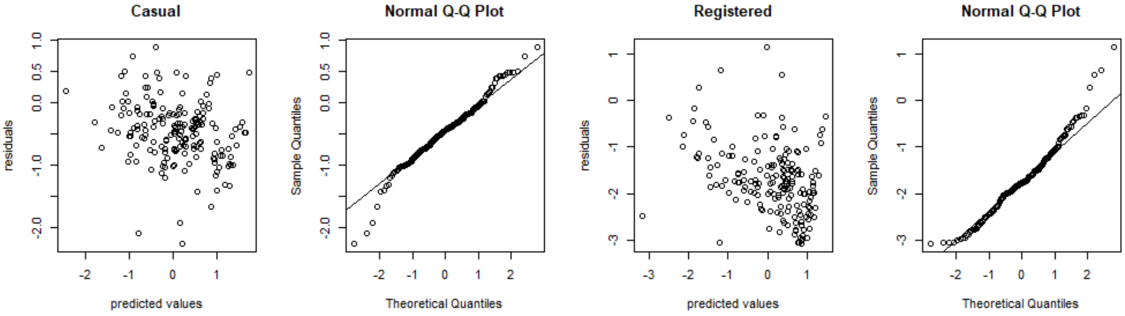

Figure 4: Residual plots of predictions for Casual (1st 2) and Registered (last 2)