CATÓLICA
CATÓLICA PORTO BUSINESS SCHOOL
PORTO

Master in Management
Course: Regression and Data Analysis
Professor: Pedro Duarte Silva

Regression Analysis of Database:
“Sao Paulo traffic”

Inês Amorim nº 355419069
José Miguel Teixeira nº 355419031
SangHoon Han nº 357919153

30 March 2020

Index

Index of Figures

Index of Tables

Introduction

In the context of the course Regression and Data Analysis, we pretend to develop a classical regression analysis (and models) about the database "Sao Paulo traffic". The database was created with records of behaviour of the urban traffic of the city of Sao Paulo in Brazil from December 14, 2009 to December 18, 2009 (From Monday to Friday), registered from 7 a.m. to 8 p.m. every 30 minutes.

Having the goal of applying the knowledge gained during this course, this report will contain insights about the regression models we used, how we select them and the analysis we did, such as autocorrelation and heteroscedasticity.

Context

Sao Paulo is a big city in Brazil, with over 12 million inhabitants, and therefore, as you can imagine, traffic is a big issue for the city, especially in rush hours, because not only creates enormous traffic jams but also pollutes the environment. Therefore, it is of great importance to know what affects more the traffic.

In the city, since 1995, there is a system of car rotation that restricts the cars on the use of the roads (for example, cars with a license plate ended by number 1 or 2 cannot circulate on Mondays or with final number 3 or 4 on Tuesdays, and so on).

Database and variables

The database "Sao Paulo traffic" gives us information regarding the slowness in traffic (continuous dependent variable) of the city, with 135 observations and 17 explanatory variables/attributes, that later on we will see which ones are relevant and if indeed influence the dependent variable. The explanatory variables are:

1. Hour (Coded, which means that 1 refers to 7 a.m., 2 refers to 7.30 a.m and so on until 27 that refers to 8 p.m.);
2. Immobilized bus;
3. Broken Truck;
4. Vehicle excess (binary variable: yes=1; no=0);
5. Accident victim;
6. Running over;
7. Fire Vehicles (binary variable: yes=1; no=0);
8. Occurrence involving freight (binary variable: yes=1; no=0);

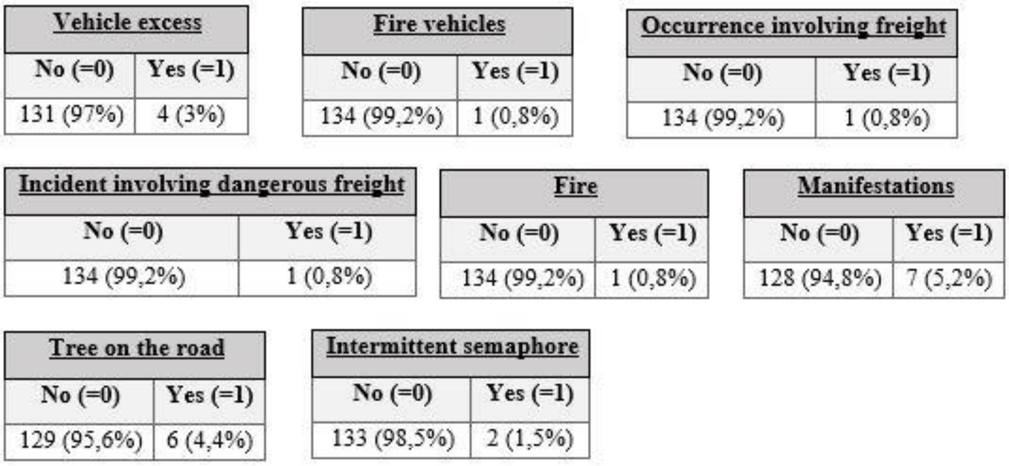
9. Incident involving dangerous freight (binary variable: yes=1; no=0);
10. Lack of electricity;
11. Fire (binary variable: yes=1; no=0);
12. Point of flooding;
13. Manifestations (binary variable: yes=1; no=0);
14. Defect in the network of trolleybuses;
15. Tree on the road (binary variable: yes=1; no=0);
16. Semaphore off;
17. Intermittent Semaphore (binary variable: yes=1; no=0).

To start we did a descriptive statistical analysis of the variables that are not binary, as seen in Table 1, and frequency analysis tables for the binary variables (Figure 1).

Variables	**Min**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**
Hour (Coded)	1	7	14	14	21	27
Immobilized bus	0	0	0	0.3407	1	4
Broken truck	0	0	1	0.8741	1	5
Accident victim	0	0	0	0.4222	1	3
Running over	0	0	0	0.1185	0	2
Lack of electricity	0	0	0	0.1185	0	4
Point of flooding	0	0	0	0.1185	0	7
Defect in the network of trolleybuses	0	0	0	0.2296	0	8
Semaphore off	0	0	0	0.1259	0	4

Table 1: Descriptive statistics of non-binary variables

According to table 1, both the mean and the median of the hour (coded) is 14, which represents 1.30 p.m. of each day. We can also state that most of the variables have the first quartile (1Q), the median and in some cases also the third quartile (3Q) equal to zero. For example, the variables point of flooding and defect in the network of trolleybuses reached a maximum of 7 and 8 observations, respectively.



Vehicle excess	
No (=0)	Yes (=1)
131 (97%)	4 (3%)

Fire vehicles	
No (=0)	Yes (=1)
134 (99,2%)	1 (0,8%)

Occurrence involving freight	
No (=0)	Yes (=1)
134 (99,2%)	1 (0,8%)

Incident involving dangerous freight	
No (=0)	Yes (=1)
134 (99,2%)	1 (0,8%)

Fire	
No (=0)	Yes (=1)
134 (99,2%)	1 (0,8%)

Manifestations	
No (=0)	Yes (=1)
128 (94,8%)	7 (5,2%)

Tree on the road	
No (=0)	Yes (=1)
129 (95,6%)	6 (4,4%)

Intermittent semaphore	
No (=0)	Yes (=1)
133 (98,5%)	2 (1,5%)

Figure 1: Frequency distribution of the binary variables

According to figure 1, we can see that situations like fires and fire vehicles on the roads and incidents involving freight and dangerous freight (like oil or gas) are very rare (happened less than 1%). From all the binary variables, the most "common" to occur were manifestations, but it's still a small number (5,2%).

Important Note: In order to analyse our data in the best way possible, we transformed the variable "Hour Coded" into a binary variable, allowing us to not treat the model as a time series anymore. We did an analysis with a plot to observe from what time there was a "turning point" and the traffic slowness started to rise. The variable "Hour Coded" is now split into regular hour (less traffic slowness hours) and rush hour (the hours with the most traffic slowness), as seen in figure 2. This way it was possible to analyse the true impact of the rush hours (with normally more people riding their cars) and regular hours in the slowness of the traffic, allowing us to have one more explanatory variable, which is no more continuous.

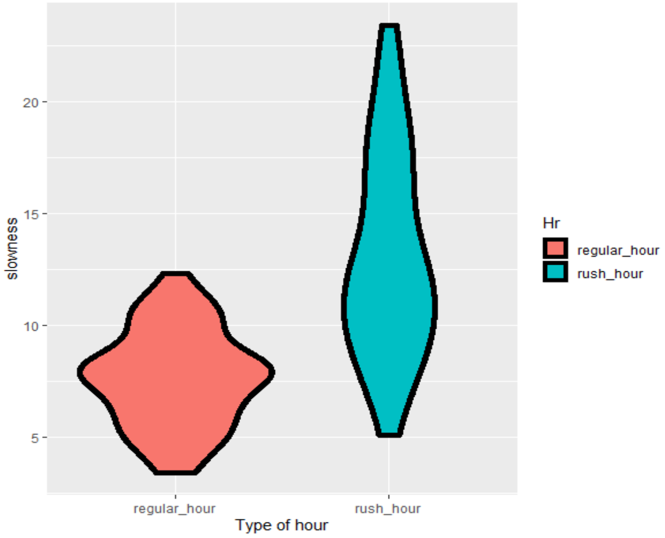


Figure 2: Slowness in traffic per type of hour

Classical Regression Analysis

Model Development

The main goal of this project is to know which of the 17 variables truly affect the slowness in Sao Paulo traffic (our dependent variable), and to do so we are going to develop regression models that can possibly explain the behaviour of the variable, having a search strategy for the (multiple) linear regression model of general to specific, because we will start with all the variables (most complete model) and then we drop the ones that are not relevant or significant.

The initial linear regression model (Model 1) is:

Slowness in traffic = β0 + β1 Hour coded + β2 Immobilized bus + β3 Broken truck + β4 Vehicle excess + β5 Accident victim + β6 Running over + β7 Fire vehicles + β8 Occurrence involving freight + β9 Incident involving dangerous freight + β10 Lack of electricity + β11 Fire + β12 Point of flooding + β13 Manifestations + β14 Defect in the network of trolleybuses + β15 Tree on the road + β16 Semaphore off + β17 Intermittent semaphore (**MODEL 1)**

After analysing the first model, we realized that there were a big number of variables that were not statistically significant. The p-value for each of these independent variable tested the null hypothesis that the variable has no correlation with the dependent variable. There wasn't a true correlation, so there was no association between the changes in the independent variable and the shifts in the dependent variable. In other words, there was insufficient evidence to conclude that there was an effect at the population level. In sum, we had to remove these variables, generating a reduced model of the model 1:

Slowness in traffic = β0 + β1 Hour coded + β2 Fire vehicles + β3 Lack of electricity + β4 Point of flooding + β5 Manifestations + β6 Defect in the network of trolleybuses + β7 Semaphore off **(MODEL 2)**

The adjusted R squared of this model has increased and the AIC has decreased significantly comparing to the previous one which shows that we are in the presence of a better model in the overall. We have chosen to analyse the adjusted R squared instead of the R squared, once the R-squared supposes that every independent variable in the model explains the variation in the dependent variable and cannot verify whether the coefficient ballpark figure and its predictions are prejudiced. So, the adjusted R-squared is a modified version of R-squared for the number of predictors in a model.

Posteriorly, we still got some variables that were not significant to the model, hereupon they had to be removed, giving rise to a new model:

Slowness in traffic = β0 + β1 Hour coded + β2 Lack of electricity + β3 Point of flooding + β4 Defect in the network of trolleybuses **(MODEL 3)**

The adjusted R has decreased just slightly and the Akaike information criterion has increase just a little too, so we believe that the three variables we have removed were not significant enough to explain our model. In order to conclude the refinement of our model by dropping variables, we removed just one more variable, because we thought it was not explanatory enough, giving rise to a new reduced model:

Slowness in traffic = β0 + β1 Hour coded + β2 Lack of electricity + β3 Point of flooding **(MODEL 4)**

The impact of this change on the adjusted R squared and in the AIC was minimal, in that order we will proceed analysing the model 4 as being our final reduced model.

However, before that, we wanted to analyse what would be the impact of removing the variables Lack of electricity and Point of flooding separately so that we can observe the impact of these too and their statistical significance to the model. We will have too more models:

Slowness in traffic = β0 + β1 Hour coded + β2 Point of flooding **(MODEL 5)**

Slowness in traffic = β0 + β1 Hour coded **(MODEL 6)**

In both cases the adjusted R squared, and the AIC changed a lot for the worse, showing that these two variables are very important for the explanation of the behaviour of the dependent variable and are statistically significant.

The model we have chosen to proceed with the validation is MODEL 4.

Normality

By opting for the regression model 4, it was necessary to ascertain if this one was satisfactory comparing to the complete one (model 1). For that, we applied the ANOVA test and considered the following two hypotheses: H0: Reduced model **(Model 4)** it's satisfactory comparing to the complete model **(Model 1);** H1**:** Otherwise.

Analysing the data and by adding the 10 variables to the reduced model, it shows that the model doesn't improve significantly, once the p-value is very large (0.4844) which is way bigger than 5%. In conclusion we won't reject H0 and we will assume that Model 4 is satisfactory.

Multicollinearity

Problems of multicollinearity occur when some explanatory variables of the regression model are linearly associated, which means that several different variables represent the same or very similar information, and we cannot know for sure which variable has a true impact in the dependent variable (which is slowness in traffic in our case).

By analysing the correlation coefficients of the integer variables of model 4 (Lack of electricity and Point of flooding), that is 0.3341486, we can state that although the variables are a bit correlated (see Figure 3), this does not represent a problem for the model because they are not excessive correlated (the problem of multicollinearity would happen if the correlation coefficient we're close to 1 or -1 in case of a negative correlation).

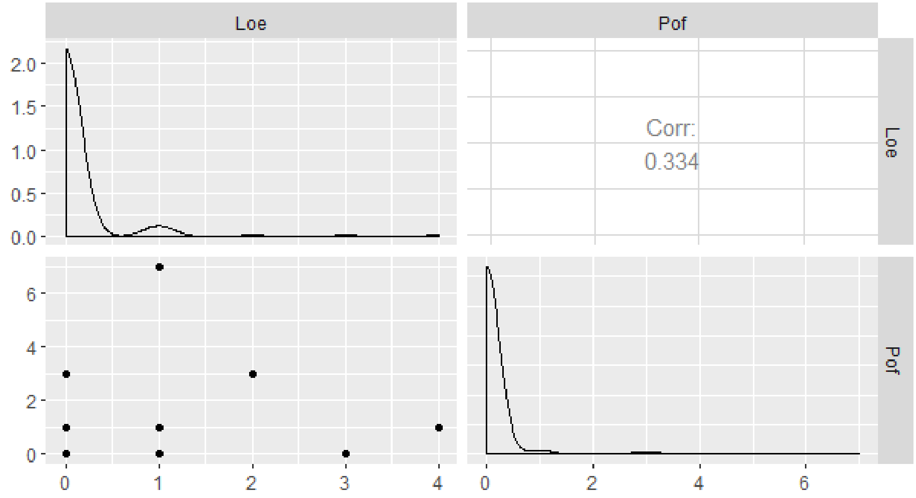


Figure 3: Correlation between the integer variables of model 4

As an informal measure to confirm our diagnostic of lack of problems of multicollinearity, we also analyse the variance inflation factor (VIF) for the 3 variables of model 4 (Hour, Lack of electricity and Point of flooding) and the values are 1.061005, 1.152792 and 1.144501, respectively. None of these VIF values are close or above 10, so we can confirm that there are no problems with multicollinearity.

Autocorrelation

OLS (ordinary least squares) regression assumes the independence of errors and observations. However, in the estimation of a time-series data regressions, it's very unlikely that the errors are independent, and that's when problems of autocorrelation can happen.

To analyse problems of autocorrelation we can use the tests of Breusch-Godfrey and Durbin-Watson, where the null hypothesis is that the errors of the regression are not correlated.

However, we haven’t done any of these tests because our observations are not a time series or cross-section data since we transformed the variable “Hour” in a factor.

Heteroscedasticity

Heteroscedasticity, which is the opposite of homoscedasticity, occurs when the variance for all the observations in a data set is not the same. OLS regression assumes homoscedasticity, therefore the variance of the errors is constant [var (y_i)=var (e_i)=σ^2]. When this assumption is violated or, in other words, when we have heteroscedasticity, we can have problems analyzing and interpret the results of our data and regression model.

In order to detect if we have this problem, we can do a residuals plot (graphical diagnostic) to see their dispersion and the Breusch-Pagan test.

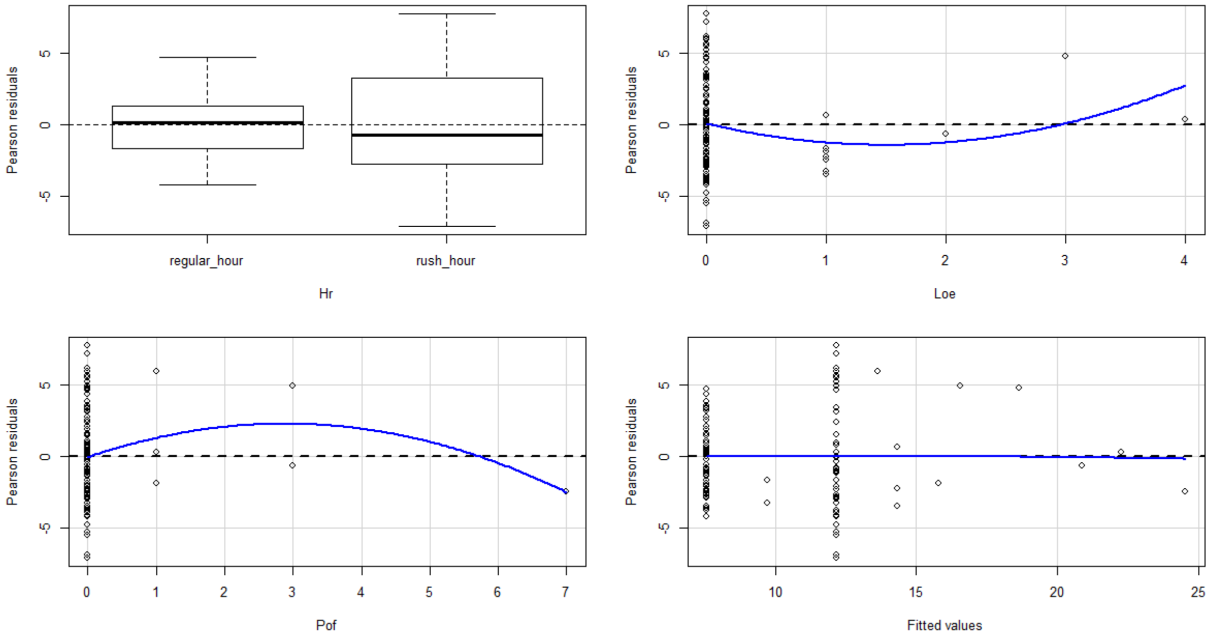


Figure 4: Residuals graph

According to figure 4, there seems to be a pattern of heteroscedasticity, especially in the variable “Lack of electricity” (Loe), but to confirm this we did a Breusch-Pagan test, where the null hypothesis is the constant variance of errors/residuals (Homoscedasticity), and the alternative hypothesis is the opposite (heteroscedasticity). With a significance level of 5%, if p-value < 0.05, we reject the null hypothesis.

In this case, the p-value is equal to 4.992e-05 (that corresponds to 0.033636), being lower than 0.05, so we reject H0. Therefore, we will have to deal with problems of heteroscedasticity. To overcome this issue, regression with robust standard errors (White) or weighted least squares are a possibility. However, we filtered the data by removing the values of when lack of electricity and point of flooding were zero, and then we did a “new” linear regression model with the variables hour and the filtered data of lack of electricity and point of flooding. The p-value of the Breusch-Pagan test is now 0.1472, which is bigger than 5% (level of significance), so we accept H0, meaning that we have now a homoscedastic model.

Outliers

We should consider the characteristics of our data when deciding on an outlier. First, our data is large when each parameter of Loe and Poe variables are zero, so the data is focused on zero points. For this reason, it’s not easy to judge outliers by the hat value and the cook’s distance. Therefore, externally studentized residual is the best option to make decisions regarding the existence of outliers, because this parameter is not only influenced largely by the leverage effect but also can be used for T-test. We did an outlier’s test and the result was the 106th row of the data with a p-value equal to 0.0090292, so we rejected H0 (H0 means that our data doesn’t have outliers), so the data of Model 4 has outliers (as seen on figure 5)**.**

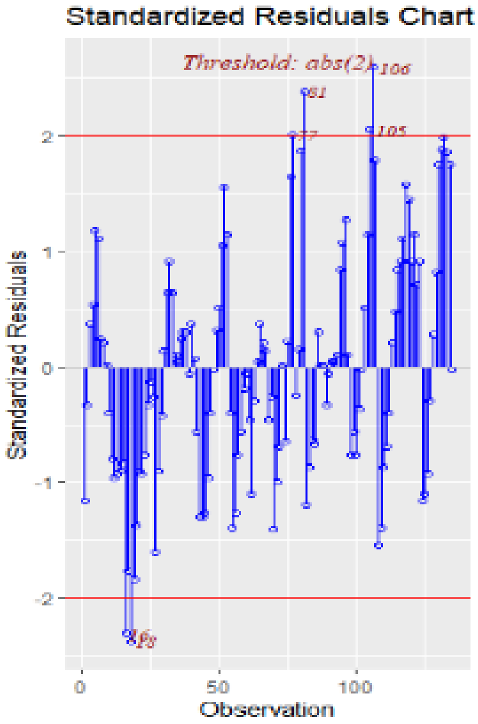


Figure 5: Outliers

The linear regression model was restructured by deleting data with outliers. Then we performed the outliers test, and the conclusion was that the new data of **Model 4** didn’t showed so much signals of having outliers once the new p-value was 0.020514, higher than 1%.

Prediction Final Model

Our final model is the Model 4 without outliers and without heteroscedasticity (mod4out). Hence, we present the summary statistics of the dependent variable, now considering the Model 4 regression maintaining the OLS estimates.

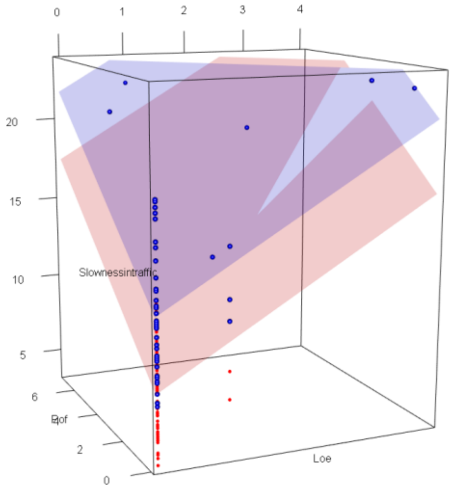
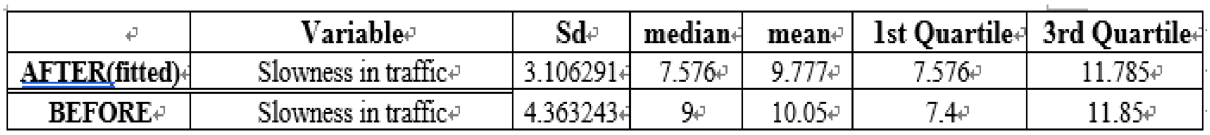


Figure 6: 3D graph - red point plotted by regular hour and blue point plotted by rush hour

In figure 6, the red plane equation represents "**Slowness in traffic** = β0 +β1*0(Regular_hour) β2 Lack of electricity + β3 Point of flooding" to focus our data on "Regular hour", and the blue plane equation is that "**Slowness in traffic** = β0 + β1*1(rush_hour) + β2 Lack of electricity + β3 Point of flooding" to focus our data on "Rush Hour".

Then we saw the data by each observation, after fitting, as seen in table 2.



	Variable	Sd	median	mean	1st Quartile	3rd Quartile
AFTER(fitted)	Slowness in traffic	3.106291	7.576	9.777	7.576	11.785
BEFORE	Slowness in traffic	4.363243	9	10.05	7.4	11.85

Table 2: Descriptive statistics of the final model 4

Finally, we created an object that contains the fitted value and upper and lower levels with a 95% confidence interval (figure 7) and the "Pdic" object that contains the fitted value and upper and lower levels with a prediction interval of 95% (figure 8). We can see that prediction interval graph shows larger upper and lower lines once the prediction interval used higher variation values than the confidence interval graph.

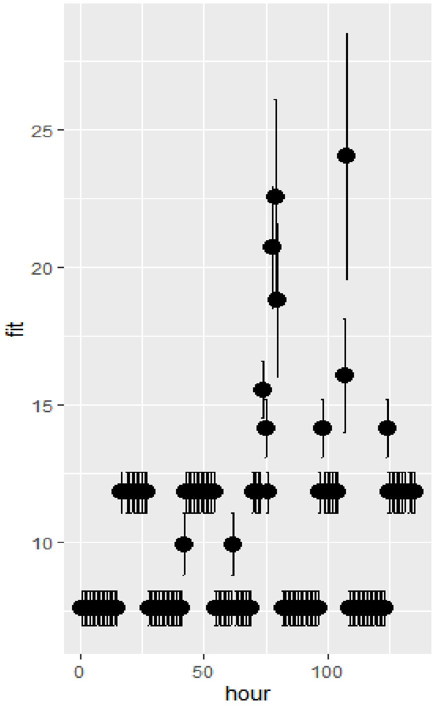


Figure 7: Confidence interval graph

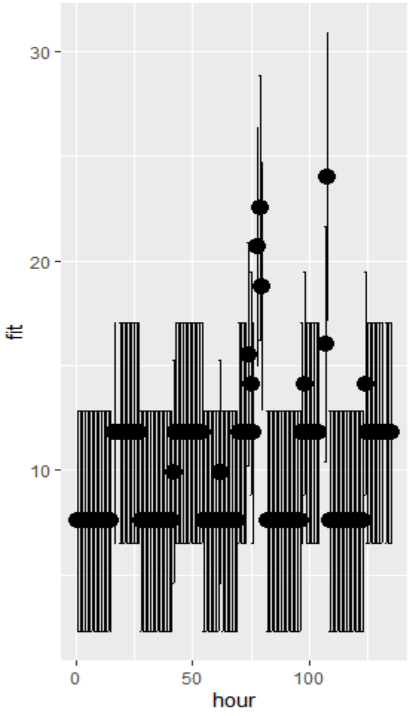


Figure 8: Prediction interval graph

Conclusion

Considering the study performed we can conclude that there are 3 major explanatory variables that are relevant for the behaviour of the depend variable Slowness in traffic. Those variables are the Hour Coded (which has been divided into Regular hours and Rush hours), Lack of Electricity and Point of Flooding.

We believe that this is our best linear regression model after making the proper needed tests and analyses to the variance, multicollinearity, heteroscedasticity and outliers, such as doing the due corrections after those.

In order to finalize, we would like to highlight that this assignment was very important to help develop our R skills and the knowledge regarding the linear model regression.