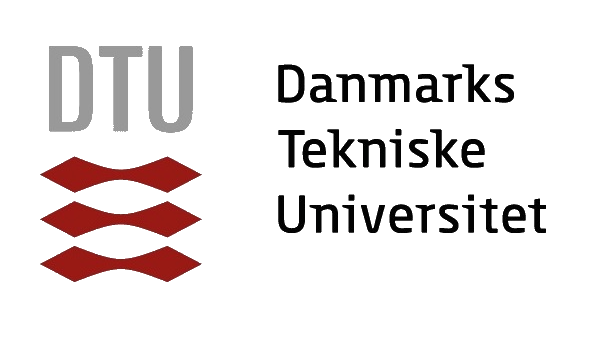
**42178**

Transport system analysis – demand and planning

Portfolio 1



**s212156 -** Irene Gómez Sánchez

**s212904 -** Zoltán György Varga

**s213470 -** Jakob Ryttergaard Poulsen

Contenido

[Exercise 1 2](#_Toc82802290)

[Reading data and libraries 2](#_Toc82802291)

[Section A: descriptive statistics 2](#_Toc82802292)

[Section B: linear regression 3](#_Toc82802293)

[Section C: new variables 5](#_Toc82802294)

[Section D: comparison 5](#_Toc82802295)

[ANOVA test 5](#_Toc82802296)

[F test 6](#_Toc82802297)

[Section E: elasticities 6](#_Toc82802298)

[Section F: predictions 7](#_Toc82802299)

[Section G: uncertainties 8](#_Toc82802300)

[Exercise 2 8](#_Toc82802301)

[Reading data and libraries 8](#_Toc82802302)

[Section A: description statistics 9](#_Toc82802303)

[Section B: first Poisson regression 9](#_Toc82802304)

[Section C: first Poisson regression 10](#_Toc82802305)

[Section D: Elasticities 11](#_Toc82802306)

[Section E: predictions 12](#_Toc82802307)

[Section F: uncertainties 12](#_Toc82802308)

[Exercise 3 13](#_Toc82802309)

[City selection 13](#_Toc82802310)

[Aspects to take into account 13](#_Toc82802311)

[The main problem in Madrid and solutions 15](#_Toc82802312)

[Value of Time 15](#_Toc82802313)

[Comfort of the public transport system 16](#_Toc82802314)

[Decrease the pollution 16](#_Toc82802315)

[References 17](#_Toc82802316)

[Appendixes 18](#_Toc82802317)

[Appendix A: initial AOV 18](#_Toc82802318)

[Appendix B: model b 19](#_Toc82802319)

[Appendix C: model b updated 20](#_Toc82802320)

[Appendix D: model b\_1 21](#_Toc82802321)

[Appendix E: model km\_per\_car\_mode 22](#_Toc82802322)

[Appendix F: model model\_ln\_b 23](#_Toc82802323)

[Appendix G: model\_ln\_b updated 24](#_Toc82802324)

[Appendix H: model c 25](#_Toc82802325)

[Appendix I: elasticities for linear models 26](#_Toc82802326)

[Appendix J: predictions in linear models. 27](#_Toc82802327)

[Appendix K: model poisson\_extended 28](#_Toc82802328)

[Appendix L: model poisson\_simple\_alr 29](#_Toc82802329)

# Exercise 1

## Section A: descriptive statistics

***Make descriptive statistics, e.g. a table with summary statistics (min, mean, std. dev., and max) and a correlation matrix, for the variables that you are going to use and discuss the results (read the whole exercise first so that you know which variables you have to use).***

Table

Description automatically generated

We note some observations about the data-set. At most, 25% of households have more than one car, as it is outside of the interquartile-range. At least 50% of houesholds have no children, some households are reporting 0 income, and the annual vehicle kilometers have not been observed lower than 553 km. We see that the annual vehicle kilometers driven has a notably higher mean value, than the median, meaning that the 50% of households that drives the most, are responsible for more than half of the driven kilometers. The annual vehicle kilometers driven might not be normally distributed, or linearly predictable.

And the correlation matrix would be:

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

We note variables that are positively correlated with VKMS: **Children, Car, HHIncome, RoadDens,** and **Worker.**

Two variables are negatively correlated with VKMS: **Elderly** and **TransSup.**

Besides, we note the correlations between the explanatory variables – **Worker** is positively correlated with **HHIncome** and **Children** and negatively correlated with **Elderly** – we note the probability of multicollinearity when including all variables.

We can also run an analysis of variance for all the variables.

A screenshot of a computer

Description automatically generated with low confidence

We note that there is evidence for all variables except for **City** and **CityShape**, having an effect on the response-variable VKMS, statistically significant at the 0.001 level.

## Section B: linear regression

***Estimate a linear regression explaining the number of vehicle km either in total or per vehicle using relevant socio-economic variables as explanatory variables. Discuss the results.***

## **All Variables**

First we construct a linear regression model, using all the socio-economic variables that showed correlation with VKMS in the correlation-matrix above, and statistical significance in the ANOVA-test. The model is of the following form:

Where Elderly and Worker are binary categorical variables

We see from the printout below that the range of numbers of children is a set from 0:5.

Table

Description automatically generated

Paying attention to the coefficients regarding **Children**, we first observe that the estimated parameters do not follow a linear progression. It seems reasonable that the more children you have, the more km you have to drive, but when you have three or more children, the effect may stop being significant.

## **Reduced Model**

We see that the estimated effect for 3, 4 and 5 children are quite similar, so to simplify the model, we decided to group number of children above three into one group.

We also see that the Elderly variable shows no statistically significant effect on the response-variable, most likely due to its high correlation with other explanatory variables, such as children, workers and income. We decide to remove it, giving us a model of the following form:

Where j is a set from 0:3, corresponding to the number of children in a household, and set equal to 3 if the number of children is higher than 3. Below we have the estimated -coefficients and .

Note: Table

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

The result is practically no change in . However, is not a good indicator when comparing two models if one of them is an extended version of the other one. In this case, adding variables can only increase the or have it remains unchanged. Usually, when the is identical, we prefer the simpler model. This “simplicity” or scarcity of variables is reflected in the F-score, in our case we see that the F-score has increased, by dropping the Elderly variable, indicating that this variable was not doing much to explain the variations of our response-variable VKMS. It is also indicated in the adjusted , which has also increased in the simpler model.

## Vehicle Kilometers per car

We can also try to change our predictor variable into annual vehicle kilometers (VKMS) per car, by dividing the two, and then constructing a model that predicts this new composite variable.

A screenshot of a computer

Description automatically generated with low confidence

In this case, and - both model metrics have been significantly reduced. Number of cars seems to be more useful to our purpose as a predictor, than as a composite of the response.

## Logarithmic Annual Vehicle Kilometers

We also have logarithmic transformations of the variables. We can try to use the natural logarithm of VKMS as our response variable, and we can include the natural logarithm of income as a predictor. First, we construct a model including all variables:

A screenshot of a computer

Description automatically generated with low confidence

We get and . It is noticeable that the R-squared is immediately higher than the purely linear models.

But it is interesting to note that the **HHIncome** variable now has a negative coefficient. In cases with multicollinearity, the coefficients should not be trusted, and due to the (obvious) high correlation between **HHIncome** and **LN\_INC,** we try to remove either one, and see that **LN\_INC** is the version of income that has the highest effect on .

We see also that two variables are not statistically significant even below the 0.1 level. Again, it is **Elderly** and **Worker**, which from the correlation-matrix seems to have a high degree of multicollinearity. We choose to keep the *Worker* variable, as it covers a wider range of people. This leaves a model of the form:

Where j is a set from 0:3, corresponding to the number of children in a household, and set equal to 3 if the number of children is higher than 3. Below we have the estimated -coefficients and .

A screenshot of a computer

Description automatically generated with medium confidence

We get and

If we consider the signs and estimated effects, all included variables are estimated as having an increasing effect on the annual vehicle kilometers driven. Intuitively, it makes sense for all of them, having children, having cars, being employed, and available roads to drive on, as seen by the road density variable, should all increase the amount of annually driven vehicle kilometers. Also, all included effects are statistically significant at the 5%-level.

## Section C: new variables

***Add the supply and city variables to the model and redo the estimation. Discuss the results.***

If we use the model for

We get and . Immediately we see an increase in , pointing towards some of the variables being able to explain additional sources of variations.

And again, we can also infer the logarithmic mode, which already excludes the main colinearity effects, *Elderly* and *HHIncome*:

Since it is our best model with the new variables, again, we show the printout:

## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.54338 0.09415 58.878 < 2e-16 \*\*\*  
## Car 0.31394 0.01655 18.973 < 2e-16 \*\*\*  
## RoadDens 1.55460 0.02502 62.137 < 2e-16 \*\*\*  
## Children\_1 0.08095 0.02076 3.899 9.82e-05 \*\*\*  
## Children\_2 0.29291 0.01931 15.173 < 2e-16 \*\*\*  
## Children\_3\_or\_more 0.38997 0.02800 13.926 < 2e-16 \*\*\*  
## Worker 0.05243 0.01927 2.720 0.00655 \*\*   
## LN\_INC 0.44925 0.01366 32.876 < 2e-16 \*\*\*  
## TransSup -0.39155 0.01552 -25.235 < 2e-16 \*\*\*  
## CityShape -0.20345 0.05198 -3.914 9.23e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.469 on 4552 degrees of freedom  
## Multiple R-squared: 0.65, Adjusted R-squared: 0.6494   
## F-statistic: 939.5 on 9 and 4552 DF, p-value: < 2.2e-16

and . Going purely from and F-score, this is our best model so far.

We see that increasing the supply of mass transit reduces the annual vehicle kilometers driven. Which makes sense intuitively, as the more mass transit supply is being offered, the more mass transit is being utilized, and the less cars will be driving on the roads. This comes with the assumption however, that any additional mass transit supply capacity, will be utilized, and will thus have a reducing effect on the annually driven vehicle kilometers. In a certain range of mass transit supply, when there is yet unfulfilled demand, this is reasonable, but at a certain point, making the population utilize mass transit is not merely a matter of increasing supply. We also see the city shape variable with a negative sign. This is a dummy variable, and the estimate shows us that for circular cities, we have less annually driven vehicle kilometers than in narrow cities. Which makes sense intuitively, as on average there will be less distance to cover between two random points in a circular layout, than in a narrow layout.

## Section D: comparison

***Compare the model from b) with the model from c). What model do you prefer and why?***

### ANOVA test

We can run an analysis of variance between the two logarithmic models:

anova(model\_ln\_b, model\_ln\_c)

## Analysis of Variance Table  
##   
## Model 1: LN\_VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more + Car +   
## LN\_INC + RoadDens + Worker  
## Model 2: LN\_VKMS ~ Car + RoadDens + Children\_1 + Children\_2 + Children\_3\_or\_more +   
## Worker + LN\_INC + TransSup + CityShape  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4554 1142.2   
## 2 4552 1001.4 2 140.84 320.1 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Thanks to this comparative analysis of variance between both models, we can conclude that Model\_ln\_c is better than Model\_ln\_b:

1. Model\_ln\_c is significantly different from Model\_ln\_b, since pvalue is <0.05.
2. Model\_ln\_c’s RSS is quite lower than Model\_ln\_b’s. This means that Model\_ln\_c manages to decrease the unexplained variance.

### F test

var.test(model\_ln\_b, model\_ln\_c, alternative = "two.sided")

##   
## F test to compare two variances  
##   
## data: model\_ln\_b and model\_ln\_c  
## F = 1.1401, num df = 4554, denom df = 4552, p-value = 9.739e-06  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 1.075783 1.208344  
## sample estimates:  
## ratio of variances   
## 1.140139

This test could be useful to compare two models if one of them is an extended version of the other one. In this test, the hypothesis contrast is the following:

1. *H0: Variances are equal, so the ratio of variances is 1*
2. *H1: Variances are unequal , so the ratio of variances is different from 1*

Results show that the difference between both models are statistically significant at 0.05.

The signs for both models are also consistent with intuitive logic, what we expect, and between eachother.

Taking into account both tests, we assume that our best model is Model\_ln\_c

## Section E: elasticities

***Calculate the average elasticity of VKMS with respect to household income and mass transit supply. Interpret the elasticity that you find.***

Firstly, we studied the interpretation of elasticities for linear models, which are in Appendix I: elasticities for linear models. However, as we have established earlier, the logarithmic model does a better job of predicting the *annual vehicle kilometers per household*. So we are mainly interested in the elasticities for this model. For the estimation of elasticities we refer to the following table from [1]

Table

Description automatically generated

Income:

For a log-log model, the elasticity is equal to the coefficient for the variable in the model.

Since the elasticity is 0.45, then a change of 10% in the income leads to an expected change of in the number of kms.

Then, for mass transit supply, we have a log-linear functional form, so we take the coefficient, multiplied with the transit-supply.

Since we are multiplying with a vector of data-points(TransitSupply), the resulting elasticity is a vector of elasticities, fitted for each individual. The average of this vector, gives us an overall, average elasticity.

Since the average elasticity is -0.35, then an increase of 10% in the transit supply mass leads to an expected change of in the total number of annually driven kilometers.

## Section F: predictions

***Calculate the effect on the driving across the cities in a future scenario where all cities are expected to increase their mass transit supply with 50% while income is expected to increase by 10% (all remaining variables are assumed to be unchanged).***

We use the previously calculated elasticities to construct this scenario, from the base data.

Table

Description automatically generated

The average change is equal to

So we predict that the given scenario would reduce the average annually driven vehicle kilometers by 2005.7 kms.

## Section G: uncertainties

***Assuming your model to be correct, briefly discuss uncertainties in your forecast in f).***

There can be different sources of uncertainties in the forecast.

On the one hand, regarding the model structure, we could face overfitting problem, which is a problem that occurs when we add more parameters (or explanatory variables) than needed trying to get a better adjustment. In these cases, it is common that more data in an expanded scenario lead to a poor adjustment. However, we have tried to avoid it by making different tests which could support our decisions to keep every variable of our best model.

Besides, there could also exist a problem with non-linearity or non-normality of the model. To study that, we plot our best model:

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

We can conclude that our model meets normality and linearity assumptions.

On the other hand, concerning model input, we could be facing a risk of non-representative data, so our predictions would be different from expected if we had taken other data. However, there is no way for us to check it at this point of the analysis.

## Influential Outliers

# Chart, scatter chart Description automatically generated

Chart, scatter chart

Description automatically generated



Chart

Description automatically generated



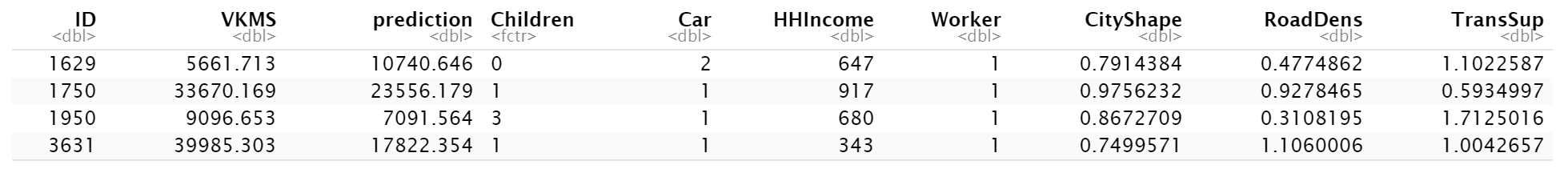
Timeline

Description automatically generated

Timeline

Description automatically generatedTimeline

Description automatically generatedTimeline

Description automatically generated

Table

Description automatically generated

# Exercise 2

## Section A: description statistics

***Make descriptive statistics, e.g. a table with summary statistics and a correlation matrix, and discuss the results, for the variables that you are going to use (read the whole exercise first to decide which variables are relevant in the descriptive statistics).***

## min Q1 median Q3 max mean sd n  
## ln(population) 8.85 10.1875 10.96 11.41 13.23 10.8533 0.9703 960  
## unemployment rate 2.30 4.2000 5.10 6.10 13.40 5.2497 1.5815 960  
## speed limit 65 0.00 0.0000 1.00 1.00 1.00 0.5490 0.4979 960  
## graduated driver license law 0.00 0.0000 0.00 1.00 1.00 0.4500 0.4978 960  
## Blood alcohol < 0.08 law 0.00 0.0000 0.00 1.00 1.00 0.4781 0.4998 960  
## Secondary seat belt law 0.00 1.0000 1.00 1.00 1.00 0.9052 0.2931 960  
## Primary seat belt law 0.00 0.0000 0.00 1.00 1.00 0.2771 0.4478 960  
## Zero tolerance alcohol law 0.00 0.0000 1.00 1.00 1.00 0.7104 0.4538 960  
## License revocation law 0.00 0.0000 1.00 1.00 1.00 0.5490 0.4979 960

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | lnpop | unempl | sp65 | gdl | bac08 | beltsc | beltpr | zerotol | alr |
| lnpop | 1.000 | 0.251 | 0.035 | 0.101 | 0.041 | 0.110 | 0.328 | 0.166 | 0.062 |
| unempl | 0.251 | 1.000 | -0.131 | -0.050 | 0.020 | -0.094 | 0.154 | -0.212 | -0.009 |
| sp65 | 0.035 | -0.131 | 1.000 | 0.530 | 0.486 | 0.114 | 0.168 | 0.543 | 0.445 |
| gdl | 0.101 | -0.050 | 0.530 | 1.00 | 0.526 | 0.150 | 0.296 | 0.545 | 0.290 |
| bac08 | 0.041 | 0.0196 | 0.486 | 0.526 | 1.000 | 0.061 | 0.241 | 0.473 | 0.591 |
| beltsc | 0.110 | -0.094 | 0.114 | 0.150 | 0.061 | 1.000 | 0.041 | 0.248 | 0.029 |
| beltpr | 0.328 | 0.154 | 0.168 | 0.296 | 0.241 | 0.041 | 1.000 | 0.216 | 0.136 |
| zerotol | 0.166 | -0.213 | 0.543 | 0.545 | 0.473 | 0.248 | 0.216 | 1.000 | 0.331 |
| alr | 0.062 | -0.009 | 0.445 | 0.290 | 0.591 | 0.028 | 0.136 | 0.331 | 1.000 |

Most of the variables are binary variables, representing the implementation(or lack thereof) of certain laws related to traffic behaviour. Hence, the mean values of the variables, will tell us the percentage of states where this law is implemented. We see that the secondary seat belt law is implemented in 90% of all states, with the rest falling a bit lower.

We notice a high positive correlation between the variables representing these regulatory actions aimed towards limiting reckless driving behavior.

## Section B: first Poisson regression

***Table 3 shows the estimation results from a Poisson regression with the number of accidents as dependent variable and using lnPop, unempl and sp65 as explanatory variables. Comment on the results.***

Model 1 estimation The first value that we look at, is the adjusted . The formula for is:

1- (-3124.2/-5026.9)

## [1] 0.3785037

And, similarly to adjusted , it represents the amount of variation that is explained by the model (Final LL), compared to a model with no parameters (Null LL). So the explanatory variables included in the model *log(Population), Unemployment rate, Speed limit of 65 mph law* are explaining some of the variation present in our measurements of the response-variable, the number of casualties in traffic among 16 year olds. We also see the p-values reflecting the t-test for the null-hypothesis of the variables having no influence on the mean being very low, meaning there is a high degree of statistical evidence that can let us reject the null-hypothesis of the variables being insignificant.

We see that population is positively correlated with the response variable. Which makes sense, as the presence of people is an essential requirement for people to get into traffic accidents. The unemployment rate has a negative correlation with the number of accidents. Perhaps because unemployed people are driving less, or have less access to vehicles. And finally we see that a speed limit of 65 mph has a reducing effect on the number of casualties.

## Section C: first Poisson regression

***Table 4 shows the estimation results from two other Poisson regressions with the number of accidents as dependent variable and additional explanatory variables. Comment on the differences among the models in Tables 3 and 4 and argue which model you prefer, e.g. by the use of LR tests as well as by looking at the signs of parameters***

Model 2 and 3 estimations Here we see two models. One model is including all the variables listed in the table, and the other model has excluded variables *gdl, Zerotol and Alr*. First thing we notice is that the adjusted is very close for the two models.

We can evaluate the two models against eachother by the log-likelihood ratio test:

So the test statistic is distributed over degrees of freedom equal to the difference in number of parameters between the two models

teststat <- -2 \* (-3068.5 - (-3030.1))  
teststat

## [1] 76.8

We use the pchisq function to evaluate this statistic over a distribution with degrees of freedom = .

pchisq(teststat, df = 9-6, lower.tail=FALSE)

## [1] 1.490256e-16

This p-value represents the amount of evidence supporting the null-hypothesis that the simpler model is equally as good at explaining the variance in the dataset, as the model with more parameters. With a p-value of ~0, we can reject the null-hypothesis.

So it seems that some of the additional variables *gdl, Zerotol and Alr* are significant at explaining the amount of variance. The variable that jumps out as most interesting to explore is alr, as it has the highest estimated effect. If we look at the correlation matrix we also see that it is positively correlated with the other two variables *zerotol* and *gdl*. So we try to reconstruct the extended model, whose printout is in Appendix K: model poisson\_extended:

And we create the simple model, but this time with *alr* included, whose printout is in Appendix L: model poisson\_simple\_alr:

And we run the test for these two models:

lrtest(poisson\_extended, poisson\_simple\_alr)

## Likelihood ratio test  
##   
## Model 1: fatal ~ lnpop + unempl + sp65 + gdl + bac08 + beltsc + beltpr +   
## zerotol + alr  
## Model 2: fatal ~ lnpop + unempl + sp65 + bac08 + beltsc + beltpr + alr  
## #Df LogLik Df Chisq Pr(>Chisq)  
## 1 10 -3030.1   
## 2 8 -3032.3 -2 4.3267 0.1149

Now we have some evidence supporting the null-hypothesis at a significance-level of 0.10. We cannot reject the null-hypothesis that these two models are similar.

In a case like this, we would prefer the simpler model, as there is a lower chance of overfitting and collinearity, but also simply for the fact that if we can get similar results while collecting data on fewer variables, the model is easier to maintain and cheaper to construct.

## Section D: Elasticities

***Based on your preferred model, calculate the min, max and average sample elasticity of expected fatalities with respect to the two variables population and unemployment rate. Comment your results.***

First we take a look at the table which contain the maximum, minimum and the mean elasticity of the population and the unemployment.

data\_ex2$elasticity\_lnpop=poisson\_extended$coefficients['lnpop']  
data\_ex2$elasticity\_unempl=data\_ex2$unempl\*poisson\_extended$coefficients['unempl']  
  
table=matrix(c(min(data\_ex2$elasticity\_lnpop),min(data\_ex2$elasticity\_unempl),max(data\_ex2$elasticity\_lnpop),max(data\_ex2$elasticity\_unempl),mean(data\_ex2$elasticity\_lnpop),mean(data\_ex2$elasticity\_unempl)),ncol=2,byrow = TRUE)  
options("scipen"=100, "digits"=3)  
colnames(table)=c('Population','Unemployment')  
rownames(table)=c('min','max','mean')  
table=as.table(table)  
table

## Population Unemployment  
## min 0.775 -1.092  
## max 0.775 -0.187  
## mean 0.775 -0.428

The three values in the case of the population is the same because in this case we get the elasticity directly from the coefficient. The value of the elasticity represents the percentage change expected by a change in the fatalities.

The elasticity in the case of Unemployment represents the same. As we can see the difference between the maximum and the minimum elasticity is very high, using the average sample elasticity infers a high degree of uncertainty for any predictions where we use the elasticity.

## Section E: predictions

***Find the effect on the expected number of fatalities of Sp65, Bac08 and Beltpr based on your final model. Discuss the results.***

data\_ex2$marginal\_sp65=poisson\_extended$coefficients['sp65']\*exp(poisson\_extended$coefficients['sp65']\*data\_ex2$sp65)  
  
data\_ex2$marginal\_bac08=poisson\_extended$coefficients['bac08']\*exp(poisson\_extended$coefficients['bac08']\*data\_ex2$bac08)  
  
data\_ex2$marginal\_beltpr=poisson\_extended$coefficients['beltpr']\*exp(poisson\_extended$coefficients['beltpr']\*data\_ex2$beltpr)  
  
  
table=matrix(c(mean(data\_ex2$marginal\_sp65),mean(data\_ex2$marginal\_bac08),mean(data\_ex2$marginal\_beltpr)),ncol=1,byrow = TRUE)  
options("scipen"=100, "digits"=3)  
colnames(table)=c('marginal effect')  
rownames(table)=c('Speed limit 65','Blood alcohol','Seat Belt')  
table=as.table(table)  
table

## marginal effect  
## Speed limit 65 -0.309  
## Blood alcohol -0.191  
## Seat Belt -0.191

The marginal effect represent that what is the effect of a coefficient to the calculated value. As it was expected, all of the coefficients have negative effect on the number of fatality. The reason all of the variables is safety measurements.

## Section F: uncertainties

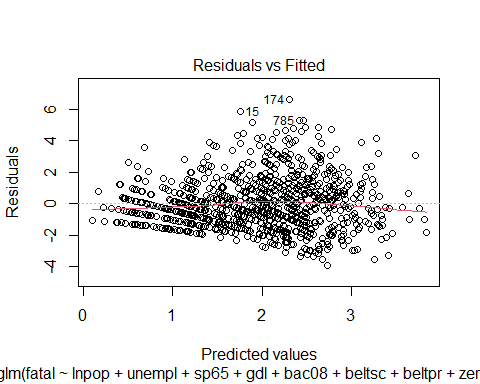
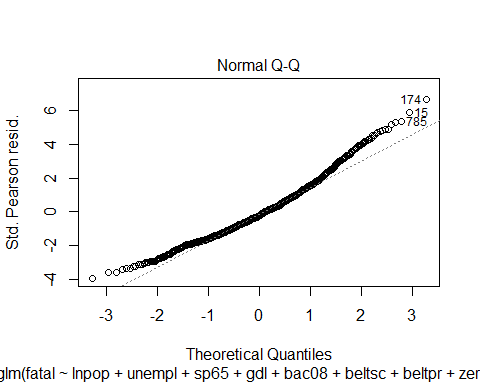
***Assuming your model to be correct, briefly discuss uncertainties in your elasticities and effect calculation in d) and e).***

As well as in the previous exercise, there can be different sources of uncertainties in the forecast.

On the one hand, regarding the model structure, we could face overfitting problem, which is a problem that occurs when we add more parameters (or explanatory variables) than needed trying to get a better adjustment. In these cases, it is common that more data in an expanded scenario lead to a poor adjustment. However, we have tried to avoid it by making different tests which could support our decisions to keep every variable of our best model.

Besides, there could also exist a problem with non-linearity or non-normality of the model. To study that, we plot our best model:

plot(poisson\_extended)

From the plots we can conclude that our model meets normality and linearity assumptions.

On the other hand, concerning model input, we could be facing a risk of non representative data, so our predictions would be different from expected if we had taken other data. However, there is no way for us to check it at this point of the analysis.

# Exercise 3

## City selection

Among the vast number of eligible cities, we decided to choose a personally known city by at least one of us: **Madrid**. The main reason is that we could address the last section of this exercise with more resources, since not only could we make suggestions but also we could discuss their feasibility according to our personal experience and knowledge about the culture. This latter is important because the culture of a city has a huge impact on its transport system development.

## Aspects to take into account

Transport system in Madrid is complex, due to the multiple means of transport that are integrated, and because it is used by a lot of people. Not only private transport system is important in this city but also public transport system. In order to know it better, it is worth taking into account its most important aspects, which are the following:

On the one hand, regarding **private transport system**, it would be determinant to know the number of cars involved, as well as the **number of people transported** by private car on average.

On the other hand, regarding **public transport system**, it would be important to know both its supply and demand. Concerning the demand, it is an important feature itself:

* **Demand:** the number of people who want to use the public transport system. In Madrid there are a lot of people who want to use (and, in fact, use) the public transport system.

With regard to the supply, there are several aspects to consider such as the number of vehicles involved, their capacity, their frequency and their coverage, to name a few. Diving deeper:

* **Supplied means of transport:** nowadays, the public transport system in Madrid (Appendix A) is made up by:
  + **533 bus routes** (218 urbans, which operate only in the city of Madrid and 315 interurban, which connect the city with the outskirts)
  + **17 metro routes** (12 conventional, 4 light metro and 1 little branch)
  + **10 commuter train routes**
  + **Public sharing bikes** (2028 bikes) and only **one bike lane** (it exists in some parts.)
* **Flexibility**: it would also be important to consider how easy it is to move through the public transport system combining different means of transport. In Madrid, since there are a lot of means of transport and each of them have different routes, the chances to combine different means are quite high. Nevertheless, with regards to bikes, the flexibility is significantly lower, since the only one bike lane that exists does not cover many places. Therefore, it is difficult to combine bikes with other means of transport.
* **Coverage**: it indicates the number of places where public transport system serves. In this sense, Madrid transport system is quite good. Usually, metro buses are an extension of either metro or train, so that everyone has at least one possibility to go to any place by public transport.
* **Connectivity**: this feature is related to how easy it is to go from one place to another place, none of them necessarily in the centre of the city. In this sense, the public transport system in Madrid is radial, which means that the connectivity between one random point and the city centre is quite good, but the connectivity between two points on the outskirts could be awful. Consequently, more private cars are used.
* **Capacity**: this feature means the number of people who can use the public transport system at the same time. This latter is important because it changes a lot in a day for example the number of people being transported at 10 am in comparison to the number of people being transported within the same day. Also, this distinction leads to other features such as the frequency of the system.
* **Supplied opportunities**: a less important feature, but not irrelevant, is how customized the price of transport could be. In this sense, Madrid allows a high customization, since there are quite different prices depending on the age and other parameters.
* **Supplied comfort**: a very significant feature of a transport system is how comfortable it is for you to travel in, because it could be the reason why a person decides to take the private car. In this sense, the comfort of the system depends a lot on the period of travelling, because the comfort in rush hours could be one tenth of the comfort in other periods of time.

In general, considering both private and public systems, it could also be a good idea to take into account the **road density**, because it would be an indicator of how much people use private and public transport.

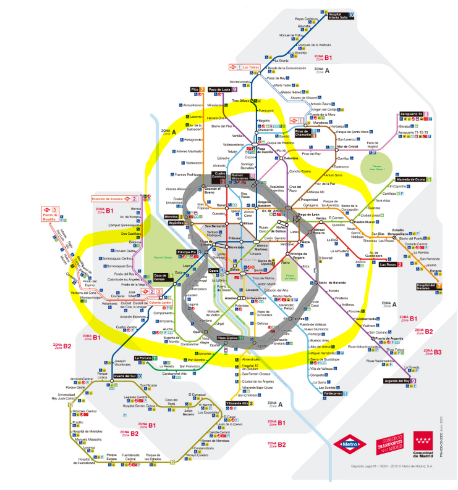
## The main problem in Madrid and solutions

The major sustainability problem which we found is the high level of pollution in Central Madrid. The main reason, which is generating this problem, is that there are more cars on the roads as should be. This is the result of the badly organized and unpopular public transport system in the city. As it was stated before, the transportation system is a radial type of system which has a central point in the central of Madrid. There are not any spherical connections between stops in the outside of the city.

We have come up with some solutions which can help to solve this problem. We can divide these solutions into three components. The first component is connected to the **Value of Time** concept. The second component is about the **comfort** of the transportation system and the vehicles. The last component is about the **direct measure** against the pollution.

Value of Time [2]

Firstly, regarding The Value of Time, it is defined by the time that a particular passenger spends traveling is considered in the value of money aggregately. We can decrease the value of time if we manage to decrease the travel time. The other way to manipulate the Value of Time is by increasing the wages, but this is not the topic of this paper.

To decrease the time, we first recommend that the policymakers should make **new circular connections on the outskirts** of the city, like an external ring. It already exists a small circular metro route, but definitely it is not enough to avoid people travelling by car, because the connections are quite poor. In this sense, the following picture shows the actual little circular ring shown in grey and the suggested ring highlighted in yellow. As a consequence of this idea, the public transport would enhance its connectivity, and, in the end, the number of cars could be also decreased. Chances of success are quite high, since the grey ring is one of the most used metro routes. Therefore, this could be understood as a high demand of metro rings, which lead to our proposal to succeed.

Another proposal is to **increase the frequency** of the public transport vehicles in the rush hours. This means the public transport operators operate their vehicles according to a Regular-interval timetable [3]. This way the number of passengers can be increased in a short time.

The last suggestion is **building parking lots** on the outskirts, so car drivers could leave their cars there and then have a **discount on the public transport** or even **make it completely free** in certain zones. Therefore, cars are stopped from entering the city centre and the congestion could be reduced, and at the same time people are economically rewarded. This solution could be especially good for Spanish people since parking is usually awful, so it both make it easier to leave your car out of the city centre and save some money from the tickets.

### Comfort of the public transport system

If we increase the public **transport frequency**, we could avoid big crowds on the vehicles. In this way we can have happier and satisfied travellers, which can make the public transport more popular. The comfort of the transport can increase if we use some restrictions on our transport system. This is that we decrease the number of car lanes and out the newly freed lanes we make bus and bicycle lanes. But the policymakers should be very careful about this method because it can lead to more congestion, since there would be less car lane, so it would be essential managing to reduce the number of cars as well. Our last proposal in this topic is a costumer program, which is further explained below, and could stimulate public transport demand.

### Decrease the pollution

The last topic was the direct way to **decrease the pollution** in the city. For that purpose, we propose restriction on the cars. It exists a project, named “Madrid Central” (Ayuntamiento de Madrid, n.d.), which also addresses this solution. It allows some cars to go through the city centre, whereas other cars can only park in specific parkings and others simply are not allowed to drive through the city centre. This can decrease the pollution in that part of the city almost to zero, taking into account that electric buses are increasingly included in the system. Next to this idea we propose that use the before mention costumer program to encourage the people to buy electric cars instead of classic petrol cars. In a long way this can decrease the pollution from the cars to zero.

# References

Ayuntamiento de Madrid, n.d. *Madrid Central. Información General.* [Online]   
Available at: https://www.madrid.es/portales/munimadrid/es/Inicio/Movilidad-y-transportes/Madrid-Central-Zona-de-Bajas-Emisiones/Informacion-general/Madrid-Central-Informacion-General/?vgnextfmt=default&vgnextoid=a67cda4581f64610VgnVCM2000001f4a900aRCRD&vgnextchannel=0 [Accessed 12 September 2021]

[1] S. Washington, F. Mannering, and P. Anastasopoulos, ‘Statistical and Econometric Methods for Transportation Data Analysis’, p. 497.

[2] J. J. Bates, ‘Value of Time’, in *International Encyclopedia of Transportation*, R. Vickerman, Ed. Oxford: Elsevier, 2021, pp. 67–71. doi: https://doi.org/10.1016/B978-0-08-102671-7.10011-9.

[3] P. Tzieropoulos, D. Emery, and J.-D. Buri, ‘Regular-interval timetables: Theoretical foundations and policy implications’, 2010.

# Appendixes

## Appendix A: initial AOV

## Df Sum Sq Mean Sq F value Pr(>F)   
## Children 1 6.534e+10 6.534e+10 3685.699 < 2e-16 \*\*\*  
## Car 1 3.969e+10 3.969e+10 2239.057 < 2e-16 \*\*\*  
## HHIncome 1 3.291e+10 3.291e+10 1856.494 < 2e-16 \*\*\*  
## ID 1 2.171e+08 2.171e+08 12.244 0.000471 \*\*\*  
## Elderly 1 3.896e+09 3.896e+09 219.750 < 2e-16 \*\*\*  
## City 1 7.971e+05 7.971e+05 0.045 0.832082   
## CityShape 1 4.938e+07 4.938e+07 2.786 0.095187 .   
## RoadDens 1 2.260e+11 2.260e+11 12748.487 < 2e-16 \*\*\*  
## TransSup 1 2.840e+10 2.840e+10 1602.143 < 2e-16 \*\*\*  
## LN\_INC 1 1.453e+10 1.453e+10 819.752 < 2e-16 \*\*\*  
## Worker 1 1.088e+09 1.088e+09 61.390 5.79e-15 \*\*\*  
## LN\_VKMS 1 2.004e+11 2.004e+11 11301.922 < 2e-16 \*\*\*  
## Residuals 4549 8.064e+10 1.773e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Appendix B: model b

## Call:  
## lm(formula = VKMS ~ Children + Car + HHIncome + Elderly + RoadDens +   
## Worker, data = data\_ex1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61823 -5582 -711 4987 32919   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.537e+04 5.905e+02 -26.027 < 2e-16 \*\*\*  
## Children1 1.654e+03 3.725e+02 4.440 9.21e-06 \*\*\*  
## Children2 5.201e+03 3.465e+02 15.008 < 2e-16 \*\*\*  
## Children3 6.547e+03 5.233e+02 12.512 < 2e-16 \*\*\*  
## Children4 6.980e+03 1.657e+03 4.213 2.57e-05 \*\*\*  
## Children5 6.745e+03 3.429e+03 1.967 0.0493 \*   
## Car 5.480e+03 2.970e+02 18.451 < 2e-16 \*\*\*  
## HHIncome 6.697e+00 3.986e-01 16.802 < 2e-16 \*\*\*  
## Elderly -7.064e+00 4.837e+02 -0.015 0.9883   
## RoadDens 2.508e+04 4.437e+02 56.511 < 2e-16 \*\*\*  
## Worker 3.334e+03 4.005e+02 8.325 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8384 on 4551 degrees of freedom  
## Multiple R-squared: 0.5385, Adjusted R-squared: 0.5375   
## F-statistic: 531 on 10 and 4551 DF, p-value: < 2.2e-16

## Appendix C: model b updated

## Call:  
## lm(formula = VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more +   
## Car + HHIncome + Elderly + RoadDens + Worker, data = data\_ex1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61818 -5584 -710 4986 32919   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.537e+04 5.904e+02 -26.033 < 2e-16 \*\*\*  
## Children\_1 1.654e+03 3.724e+02 4.441 9.16e-06 \*\*\*  
## Children\_2 5.201e+03 3.465e+02 15.012 < 2e-16 \*\*\*  
## Children\_3\_or\_more 6.585e+03 5.013e+02 13.135 < 2e-16 \*\*\*  
## Car 5.481e+03 2.969e+02 18.461 < 2e-16 \*\*\*  
## HHIncome 6.696e+00 3.985e-01 16.805 < 2e-16 \*\*\*  
## Elderly -7.550e+00 4.835e+02 -0.016 0.988   
## RoadDens 2.507e+04 4.435e+02 56.535 < 2e-16 \*\*\*  
## Worker 3.333e+03 4.003e+02 8.326 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8382 on 4553 degrees of freedom  
## Multiple R-squared: 0.5385, Adjusted R-squared: 0.5377   
## F-statistic: 664.1 on 8 and 4553 DF, p-value: < 2.2e-16

## Appendix D: model b\_1

model\_b\_1 <- lm(VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more + Car + HHIncome + RoadDens + Worker, data = data\_ex1)  
summary(model\_b\_1)

## Call:  
## lm(formula = VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more +   
## Car + HHIncome + RoadDens + Worker, data = data\_ex1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61816 -5584 -710 4986 32919   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.537e+04 5.243e+02 -29.324 < 2e-16 \*\*\*  
## Children\_1 1.655e+03 3.705e+02 4.465 8.18e-06 \*\*\*  
## Children\_2 5.201e+03 3.450e+02 15.076 < 2e-16 \*\*\*  
## Children\_3\_or\_more 6.585e+03 4.999e+02 13.174 < 2e-16 \*\*\*  
## Car 5.481e+03 2.968e+02 18.465 < 2e-16 \*\*\*  
## HHIncome 6.696e+00 3.983e-01 16.814 < 2e-16 \*\*\*  
## RoadDens 2.507e+04 4.434e+02 56.550 < 2e-16 \*\*\*  
## Worker 3.337e+03 3.248e+02 10.273 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8381 on 4554 degrees of freedom  
## Multiple R-squared: 0.5385, Adjusted R-squared: 0.5378   
## F-statistic: 759.1 on 7 and 4554 DF, p-value: < 2.2e-16

## Appendix E: model km\_per\_car\_mode

km\_per\_car\_model <- lm(VKMS/Car ~ Children\_1 + Children\_2 + Children\_3\_or\_more + HHIncome + RoadDens + Worker, data = data\_ex1)

summary(km\_per\_car\_model)

## Call:  
## lm(formula = VKMS/Car ~ Children\_1 + Children\_2 + Children\_3\_or\_more +   
## HHIncome + RoadDens + Worker, data = data\_ex1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40038 -5226 -1161 3691 34080   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6994.6156 400.8325 -17.450 < 2e-16 \*\*\*  
## Children\_1 1222.1148 351.0207 3.482 0.000503 \*\*\*  
## Children\_2 4424.7996 326.8489 13.538 < 2e-16 \*\*\*  
## Children\_3\_or\_more 4721.9165 472.5197 9.993 < 2e-16 \*\*\*  
## HHIncome 4.5153 0.3726 12.119 < 2e-16 \*\*\*  
## RoadDens 21800.4057 420.0469 51.900 < 2e-16 \*\*\*  
## Worker 2771.9391 306.7017 9.038 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7940 on 4555 degrees of freedom  
## Multiple R-squared: 0.4438, Adjusted R-squared: 0.443   
## F-statistic: 605.7 on 6 and 4555 DF, p-value: < 2.2e-16

## Appendix F: model model\_ln\_b

model\_ln\_b <- lm(LN\_VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more + Car + HHIncome + LN\_INC + Elderly + RoadDens + Worker, data = data\_ex1)

summary(model\_ln\_b)

## Call:  
## lm(formula = LN\_VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more +   
## Car + HHIncome + LN\_INC + Elderly + RoadDens + Worker, data = data\_ex1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.44532 -0.31544 0.02316 0.33010 1.77629   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.720e+00 1.184e-01 39.872 < 2e-16 \*\*\*  
## Children\_1 9.336e-02 2.226e-02 4.195 2.78e-05 \*\*\*  
## Children\_2 3.054e-01 2.071e-02 14.748 < 2e-16 \*\*\*  
## Children\_3\_or\_more 3.979e-01 2.995e-02 13.285 < 2e-16 \*\*\*  
## Car 3.245e-01 1.773e-02 18.307 < 2e-16 \*\*\*  
## HHIncome -1.028e-04 3.521e-05 -2.920 0.00351 \*\*   
## LN\_INC 5.022e-01 2.158e-02 23.268 < 2e-16 \*\*\*  
## Elderly -2.921e-02 2.893e-02 -1.010 0.31258   
## RoadDens 1.630e+00 2.648e-02 61.539 < 2e-16 \*\*\*  
## Worker 3.120e-02 2.513e-02 1.242 0.21440   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5004 on 4552 degrees of freedom  
## Multiple R-squared: 0.6016, Adjusted R-squared: 0.6008   
## F-statistic: 763.9 on 9 and 4552 DF, p-value: < 2.2e-16

## Appendix G: model\_ln\_b updated

model\_ln\_b <- lm(LN\_VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more + Car + LN\_INC + RoadDens + Worker + Elderly, data = data\_ex1)  
summary(model\_ln\_b)

## Call:  
## lm(formula = LN\_VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more +   
## Car + LN\_INC + RoadDens + Worker + Elderly, data = data\_ex1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.44079 -0.31421 0.02337 0.33217 1.78108   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.95094 0.08827 56.087 < 2e-16 \*\*\*  
## Children\_1 0.09331 0.02228 4.189 2.85e-05 \*\*\*  
## Children\_2 0.30248 0.02070 14.612 < 2e-16 \*\*\*  
## Children\_3\_or\_more 0.39778 0.02998 13.269 < 2e-16 \*\*\*  
## Car 0.31976 0.01767 18.100 < 2e-16 \*\*\*  
## LN\_INC 0.45570 0.01459 31.228 < 2e-16 \*\*\*  
## RoadDens 1.62958 0.02650 61.489 < 2e-16 \*\*\*  
## Worker 0.03599 0.02510 1.434 0.152   
## Elderly -0.02368 0.02889 -0.820 0.412   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5008 on 4553 degrees of freedom  
## Multiple R-squared: 0.6009, Adjusted R-squared: 0.6002   
## F-statistic: 856.8 on 8 and 4553 DF, p-value: < 2.2e-16

## Appendix H: model c

## model\_c <- lm(VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more + Car + HHIncome ## + RoadDens + TransSup + Worker + CityShape, data = data\_ex1)  
## summary(model\_c)

## Call:  
## lm(formula = VKMS ~ Children\_1 + Children\_2 + Children\_3\_or\_more +   
## Car + HHIncome + RoadDens + TransSup + Worker + CityShape,   
## data = data\_ex1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61472 -5250 -450 4495 29953   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7048.1929 930.9678 -7.571 4.46e-14 \*\*\*  
## Children\_1 1441.3309 353.6671 4.075 4.67e-05 \*\*\*  
## Children\_2 5030.0802 329.2421 15.278 < 2e-16 \*\*\*  
## Children\_3\_or\_more 6439.4118 476.9862 13.500 < 2e-16 \*\*\*  
## Car 5380.9143 283.2420 18.998 < 2e-16 \*\*\*  
## HHIncome 6.6675 0.3799 17.552 < 2e-16 \*\*\*  
## RoadDens 23991.6538 426.3708 56.269 < 2e-16 \*\*\*  
## TransSup -5619.9236 264.4173 -21.254 < 2e-16 \*\*\*  
## Worker 3354.4276 309.8276 10.827 < 2e-16 \*\*\*  
## CityShape -3107.9636 886.0224 -3.508 0.000456 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7994 on 4552 degrees of freedom  
## Multiple R-squared: 0.5803, Adjusted R-squared: 0.5795   
## F-statistic: 699.3 on 9 and 4552 DF, p-value: < 2.2e-16

## Appendix I: elasticities for linear models

If we look at the linear model and calculate the elasticity for **Household Income**:

data\_ex1$prediction\_c <- predict(model\_c, data\_ex1)  
data\_ex1$elasticity\_HHIncome\_c = model\_c$coefficients["HHIncome"]\*(data\_ex1$HHIncome/data\_ex1$prediction\_c)  
head(select(data\_ex1, c("VKMS", "prediction\_c", "HHIncome", "elasticity\_HHIncome\_c")))

## # A tibble: 6 x 4  
## VKMS prediction\_c HHIncome elasticity\_HHIncome\_c  
## <dbl> <dbl> <dbl> <dbl>  
## 1 11591. 16959. 488 0.192  
## 2 7842. 18726. 753 0.268  
## 3 10325. 17859. 623 0.233  
## 4 23139. 28343. 634 0.149  
## 5 12938. 26014. 1039 0.266  
## 6 9097. 14025. 551 0.262

mean(data\_ex1$elasticity\_HHIncome\_c)

## [1] 0.5394272

Since the average elasticity is 0.32, then a change of 10% in HH\_income leads to an expected change of in the number of kms.

Then we look at the elasticity for **Transit supply**:

data\_ex1$elasticity\_TransSup\_c = model\_c$coefficients["TransSup"]\*(data\_ex1$TransSup/data\_ex1$prediction\_c)  
head(select(data\_ex1, c("VKMS", "prediction\_c", "TransSup", "elasticity\_TransSup\_c")))

## # A tibble: 6 x 4  
## VKMS prediction\_c TransSup elasticity\_TransSup\_c  
## <dbl> <dbl> <dbl> <dbl>  
## 1 11591. 16959. 1.13 -0.375  
## 2 7842. 18726. 1.13 -0.340  
## 3 10325. 17859. 1.13 -0.356  
## 4 23139. 28343. 1.13 -0.224  
## 5 12938. 26014. 1.13 -0.244  
## 6 9097. 14025. 1.13 -0.453

mean(data\_ex1$elasticity\_TransSup\_c)

## [1] -1.190068

Since the average elasticity is -0.12, then a change of 10% in Transit supply leads to an expected change of in the number of kms.

## Appendix J: predictions in linear models.

We make our predictions using our elasticities calculated. First, for the linear model:

data\_ex1$prediction\_new\_elasticity =   
 data\_ex1$prediction\_c + data\_ex1$HHIncome\*0.10\*mean(data\_ex1$elasticity\_HHIncome\_c) +  
 data\_ex1$TransSup\*0.50\*mean(data\_ex1$elasticity\_TransSup\_c)  
head(select(data\_ex1,c("prediction\_new\_elasticity","prediction\_c")))

## # A tibble: 6 x 2  
## prediction\_new\_elasticity prediction\_c  
## <dbl> <dbl>  
## 1 16985. 16959.  
## 2 18766. 18726.  
## 3 17892. 17859.  
## 4 28377. 28343.  
## 5 26069. 26014.  
## 6 14054. 14025.

mean(data\_ex1$prediction\_new\_elasticity - data\_ex1$prediction\_c)

## [1] 32.94698

On average, the linear model predicts a 19.7 km increase in annual vehicle kilometers per household in this scenario.

## Appendix K: model poisson\_extended

summary(poisson\_extended)

## ## Call:  
## glm(formula = fatal ~ lnpop + unempl + sp65 + gdl + bac08 + beltsc +   
## beltpr + zerotol + alr, family = poisson(link = "log"), data = data\_ex2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.5435 -1.3324 -0.2366 0.8624 5.2923   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.65905 0.15539 -36.418 < 2e-16 \*\*\*  
## lnpop 0.77523 0.01413 54.856 < 2e-16 \*\*\*  
## unempl -0.08151 0.00818 -9.965 < 2e-16 \*\*\*  
## sp65 -0.37294 0.02889 -12.910 < 2e-16 \*\*\*  
## gdl -0.04579 0.03078 -1.488 0.137   
## bac08 -0.20948 0.03169 -6.610 3.84e-11 \*\*\*  
## beltsc -0.18339 0.03633 -5.048 4.47e-07 \*\*\*  
## beltpr -0.20127 0.02690 -7.483 7.28e-14 \*\*\*  
## zerotol 0.05269 0.03128 1.685 0.092 .   
## alr 0.22297 0.02723 8.188 2.66e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 6670.4 on 959 degrees of freedom  
## Residual deviance: 2676.9 on 950 degrees of freedom  
## AIC: 6080.3  
##   
## Number of Fisher Scoring iterations: 5

## Appendix L: model poisson\_simple\_alr

summary(poisson\_simple\_alr)

##   
## Call:  
## glm(formula = fatal ~ lnpop + unempl + sp65 + bac08 + beltsc +   
## beltpr + alr, family = poisson(link = "log"), data = data\_ex2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.4588 -1.3591 -0.2206 0.8570 5.2471   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.673754 0.155165 -36.566 < 2e-16 \*\*\*  
## lnpop 0.779800 0.013902 56.092 < 2e-16 \*\*\*  
## unempl -0.084852 0.007844 -10.817 < 2e-16 \*\*\*  
## sp65 -0.377735 0.025651 -14.726 < 2e-16 \*\*\*  
## bac08 -0.216825 0.029050 -7.464 8.41e-14 \*\*\*  
## beltsc -0.181892 0.035992 -5.054 4.33e-07 \*\*\*  
## beltpr -0.202643 0.026448 -7.662 1.83e-14 \*\*\*  
## alr 0.230353 0.026988 8.535 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 6670.4 on 959 degrees of freedom  
## Residual deviance: 2681.2 on 952 degrees of freedom  
## AIC: 6080.6  
##   
## Number of Fisher Scoring iterations: 5