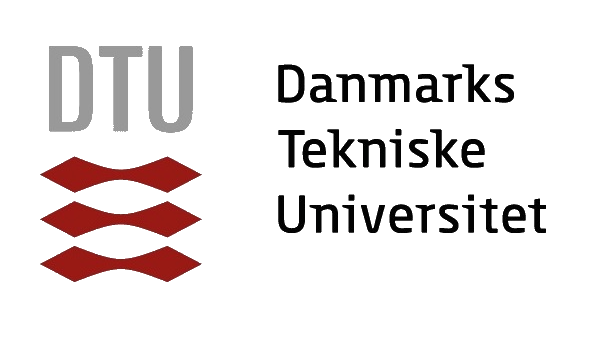
**42178**

Transport system analysis – demand and planning

Portfolio 2



**s212156 -** Irene Gómez Sánchez

**s212904 -** Zoltán György Varga

**s213470 -** Jakob Ryttergaard Poulsen

Content

[List of figures 2](#_Toc84957220)

[Exercise 1 3](#_Toc84957221)

[Section 1: descriptive statistics 3](#_Toc84957222)

[Section 2: Poisson model 4](#_Toc84957223)

[Section 3: trip generation model based 5](#_Toc84957224)

[Section 4: multinomial logit model 8](#_Toc84957225)

[Section 5: applied multinomial logit model 8](#_Toc84957226)

[Section 6: difference between Logit and Poisson 10](#_Toc84957227)

[Exercise 2 12](#_Toc84957228)

[Section 1: Distribution model 12](#_Toc84957229)

[Section 2: Distribution new model 12](#_Toc84957230)

[Section 3: Gravity model 13](#_Toc84957231)

[Section 4: Evaluate models and select the best one 13](#_Toc84957232)

[Section 5: Prediction on best model 14](#_Toc84957233)

# List of figures

[Figure 1. Exercise 1.1: Statistics descriptive 2](#_Toc84957113)

[Figure 2. Exercise 1.1: Correlation matrix 2](#_Toc84957114)

[Figure 3. Exercise 1.1: Trips distribution 3](#_Toc84957115)

[Figure 4. Exercise 1.2: Data and Poisson distribution 4](#_Toc84957116)

[Figure 5. Exercise 1.2: Comparison between empirical and theoretical Poisson distributions 4](#_Toc84957117)

[Figure 6. Exercise 1.2: expected number of individuals 4](#_Toc84957118)

[Figure 7. Exercise 1.5: Expected number of individuals 9](#_Toc84957119)

[Figure 8. Exercise 1.5: Histogram of Trips 1 9](#_Toc84957120)

[Figure 9. Exercise 1.6: Results in Poisson model 10](#_Toc84957121)

[Figure 10. Exercise 1.6: Results in Logit model 10](#_Toc84957122)

[Figure 11. Exercise 2.1: IPF procedure 11](#_Toc84957123)

[Figure 12. Exercise 2.2: IPF procedure 11](#_Toc84957124)

[Figure 13. Exercise 2.3: k and gamma values 12](#_Toc84957125)

[Figure 14. Exercise 2.3: IPF procedure 12](#_Toc84957126)

[Figure 15. RMSE formula Figure 16. Correct trip matrix for the base year 12](#_Toc84957127)

[Figure 17. Exercise 2.4: Comparison of models by RMSE 13](#_Toc84957128)

[Figure 18. Exercise 2.4: Graphical comparison of models 13](#_Toc84957129)

[Figure 19. Exercise 2.5: OiDj values Figure 20. Exercise 2.5: AB values Figure 21. Exercise 2.5. gcij values 14](#_Toc84957130)

[Figure 22. Exercise 2.5: results 14](#_Toc84957131)

# Exercise 1

## Section 1: descriptive statistics

**Analyse the data using descriptive statistics, e.g., summary statistics and a correlation matrix. What is the average number of trips using Trips and Trips1? Also plot the distribution of trips using the Trips1 variable. Comment your results.**

The results for statistics descriptive are showed in Figure 1:

Table

Description automatically generated

Figure 1. Exercise 1.1: Statistics descriptive

The average number of trips using Trips is 3.12 and using Trips1 is 2.84, and first three quartiles are the same. Therefore, both variables must be similar, and the difference between them is just the group 5+. The maximum number of trips in Trips is 16, but there cannot be a lot of individuals with many more than 5 trips, since the mean is close to Trips1, variable whose maximum value is 5. And the correlation matrix is:

*Chart, diagram, scatter chart

Description automatically generated*

Text

Description automatically generated

Regarding sizes (Figure 2), we note that explanatory/dependent variables are not highly correlated with each other, and independent variables are not highly correlated with dependent ones. However, Trips and Trips1 are highly correlated since they are quite similar.

About signs in Figure 2, for example, both Trips and Trips1 are positively correlated with Income and Cars, which makes sense. However, they are negatively correlated with Male and Age. This latter could be explained by the fact that young people need to move more and less directly than adults, who usually have cars and can travel more directly.

In Figure 3 it is confirmed what it was supposed above: there are not many observations which take more than 5 trips per day.

Chart, histogram

Description automatically generated

Figure 3. Exercise 1.1: Trips distribution



## Section 2: Poisson model

**Assume you have estimated a Poisson model with the number of trips as dependent variable based on Trips1 with the following specification for the expected number of tours:**



**where 1 is added to income to allow for ln(). We get the following results**

Tabla

Descripción generada automáticamente

**What can you say about the results and what is the expected number of trips for your sample according to the model? In addition, try to analyse the expected number of individuals with 0,1,2,3,4,5+ trips based on your model. Finally, use your estimated model to simulate the effect from a 20% increase in income on the number of trips. You can use either the income elasticity or simulation (the latter is like what you have to do for the logit model in step 5)**

We see from the table above, that all parameters are statistically significant at the 5%-level. We have a positive coefficient for income and cars, meaning both variables increase the most likely number of trips that an individual is expected to make. We have negative coefficients for age and male, meaning that an increase in age reduces the likely number of trips an individual is going to make, and males are expected to take less trips than females.

Gráfico, Histograma

Descripción generada automáticamente

Figure 4. Exercise 1.2: Data and Poisson distribution

According to the model (defined by the coefficients provided in the initial formulation), the average expected number of trips per person for the sample is 2.89. It was calculated as the mean of the predicted values from the model for all the data.

This distribution shows the most likely amount of trips people in our sample are expected to take. To calculate the expected number of individuals based on a Poisson model, we need to evaluate the probabilities of the distribution.

The poisson distribution is characterized by one parameter,

We then evaluate the probability of taking 0, 1, 2, 3, 4, and 5+ trips and multiply the probabilities with the number of people in the sample, n=1000. And we get the following results

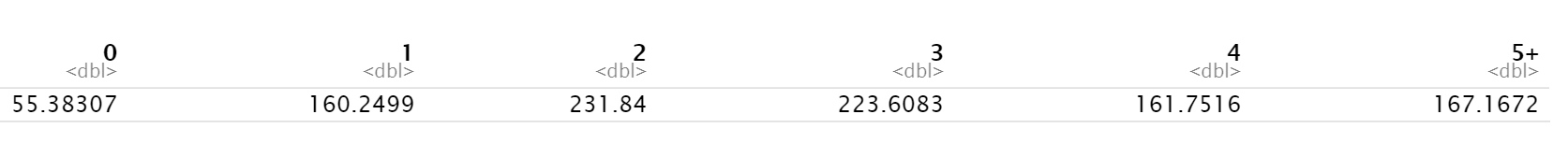


Figure 6. Exercise 1.2: expected number of individuals

Chart, bar chart

Description automatically generated

We can evaluate elasticities in the poisson-model, if we take the natural logarithm on both sides, to force the model into a linear form.

And we utilize the relations derived in this table:

Table

Description automatically generated

We see that the elasticity of income in regards to the log-likelihood of the expected number of trips is a log-log relationship, so it is given as the beta-coefficient 0,058. A 20% increase in income would lead to a 20%\*0,058=1,16% increase in the average expected number of trips.

So to simulate this situation, we would have to increase the scale-parameter by 1.16% and reevaluate the number of trips



Chart, bar chart

Description automatically generated

If we have the total number of trips estimated by the poisson model as 2.89\*1000=2890, the predicted total number of trips after a 20% increase in income would be:

So an increase in trips of 33.5

## Section 3: trip generation model based

**Specify a trip generation model based on cross-classification where you use at least two of the four variables in the Trips2020 data. You can use the quartiles of income to divide income in four intervals. Use the data to determine rates for different groups of individuals and use these rates when income is increased by 20% in a scenario. Discuss the difference between the cross-classification results and the results of the Poisson model.**

The principle behind cross-classification is the following equation:

Text

Description automatically generated

So if we know the average number of trips for a household(or person), we can multiply it with the number of households(or persons) in the population, and we will get an estimation of the total number of trips for this group.

So when we have a sample like we have, we would like to group the persons, according to the assigned values of the measured variables. We will slice age and income into four groups, based on the quantiles, seen in the descriptive statistics above, and we will group number of cars as 0, 1 and 2 or above.

When we try this, n is too small for some groups, so we try to remove the age variable, and instead only group people by income and number of cars, and we get 12 groups and the following estimations for average number of trips for each group:

Table

Description automatically generated

If we increase the income variable by 20% and sort the measurements into groups, based on the same income quartiles as used previously, then we can multiply the number of people in this group, with the average number of trips estimated for this group, as reported in the table above. Then we get the following estimation of trips

Table

Description automatically generated

And if we then subtract the measured number of trips from the newly estimated number of trips:

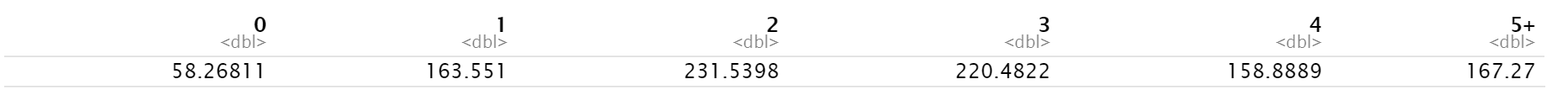
sum(cross\_classification\_2$Cross\_classification\_trips - cross\_classification\_2$Trips)

[1] 38.68826

So we would estimate an increase of 38.6 trips.

For every prototypical house-hold we have an average number of trips, and a number of households. If we assume that the probability of taking a certain number of trips follow a poisson distribution, we may set the average trips for a household, , evaluate the probabilities in a poisson distribution and multiply with , number of households.

For the entire sample, when we sum up the evaluations for every household-type, this gives us the distribution shown below:



Chart, bar chart

Description automatically generated

Limitations of the poisson-distribution: it is centered around the mean value – but we see from the empirical data, that people are much more likely to make either 0, 2 or 4 trips compared to an uneven number of trips, 1 and 3.

## Section 4: multinomial logit model

**Based on the data, suppose that you have estimated a multinomial logit model for the number of trips with the following utility specifications and that you get the following results:**

Texto, Carta

Descripción generada automáticamenteTabla

Descripción generada automáticamente

The principle behind multinomial logit models is Random Utility Maximization (RUM). We assume that a choice of number of trips can be associated with a latent amount of utility. The more utility a certain choice provides, the more likely that person is to make said choice. While the given utility for a specific person is latent and unmeasureable, we can measure people’s choices and from them estimate systematic utility for a population. We split the utility into systematic utility and a stochastic error-term .

This means that we do not have to adhere to the center of the Poisson distribution, and we can differentiate more between number of trips.

* Regarding Ks: We see that the values for the alternative-specific constants (Ks) is lower for 1, 3, and 5 trips.
* Concerning B1 and B2, the income slightly increases the utility of people to make trips and have the same effect for 2,3 and 4 tours. It increases a little bit more the utility of making 5 trips.
* About B3, having a car increases the utility of making a trip and have the same effect for making 3, 4 or 5 trips.
* Regarding B4 and B5, ages significantly reduce a little the utility of having 2, 3, 4, and 5 trips, but slightly less so in the case of 2 trips.
* About B6, the utility for males to make trips (2, 3, 4 or 5) is lower than the utility for women. It means that women are more likely to make more trips.

## Section 5: applied multinomial logit model

**Apply the estimated model in 4) to the data**

1. *Based on the estimated parameters, predict the total number of trips for the data.*

First, we created the utilities needed to calculate the probabilities. When we get the probabilities in every case, we can use the presented equation to calculate the estimation.

For that, utility functions are obtained as follows:

And, afterwards, choice probabilities can be calculated as:

When we aggregate the predictions, by multiplying the average probabilities with the sample size, we get the estimated number of people taking X amount of trips.

And the total number of trips that the model predicts is 2821.865 trips.

*b) Analyse the expected number of individuals with 0,1,2,3,4,5+ trips based on your model. Compare to the sample shares*

We can adjust the equation above, to not tell us trips, but to tell us how many people are expected to take a certain number of trips:

When we do it for number of trips being equal to = {0,1,2,3,4,5+}, we get the following six results:



Original:



Chart, bar chart, histogram

Description automatically generated

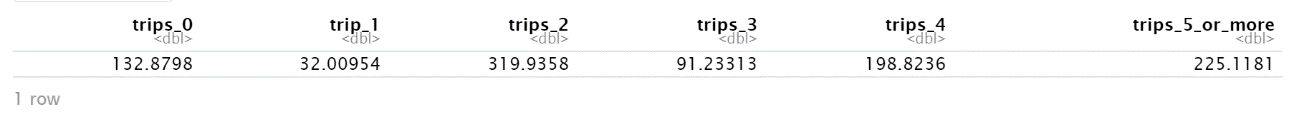
Now we have a model that more clearly reflects the dataset showing that people are less likely to take an uneven number of trips (most likely because most people return home after taking a trip).

*c) In addition, simulate the total number of trips when income is increased by 20%. Based on this simulation, calculate the elasticity of the total number of trips with respect to income.*

We created new income values which are 1.2 times higher than the original income values and recalculated the utility functions:

And, afterwards, choice probabilities are calculated:

First we evaluate how many people we now predict to make 0,1,2,3,4 and 5+ trips



And the estimated number of trips when income has been increased by 20% is 2866,466. Meaning an increase of 44,6 trips. If we use these numbers, we can calculate the elasticity of trips in regards to income by

## Section 6: difference between Logit and Poisson

*Discuss the difference* *between results from the logit and the Poisson model.*

**Poisson model:**

To calculate predictions based on a Poisson model, a Poisson distribution function is used in R. They are calculated the probabilities of making 0 to 5+ trips and then they are multiplied by the total number of people, 1000. Results are shown in the Figure X:



Figure 9. Exercise 1.6: Results in Poisson model

**Logit model:**

To calculate predicts based on a Logit model, the same procedure is used, but with a Logit distribution. Results are shown in the Figure X:



Figure 10. Exercise 1.6: Results in Logit model

We see that in this case, for trips, the Poisson model suffers from the centrality restriction of the distribution. It is not possible to differentiate between different choices, and to put lower probabilities on some choices that are near the center.

# Exercise 2

## Section 1: Distribution model

**Assume a model with independence among origins and destinations, i.e. 𝑇𝑖j=𝐴𝑖𝐵𝑗𝑂𝑖𝐷𝑗. Calculate the base matrix, 𝑇ijB, using this model. Do you think this is a useful model assuming independence?**

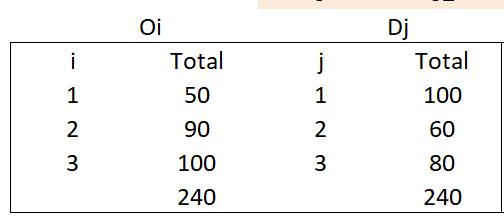
We are given marginal totals and no initial solution, but an assumption of independence among origins and destinations(this is rarely the case). One way of fitting the OD matrix under assumption of independence is the population-count method:

We can enter the proportions of and as:

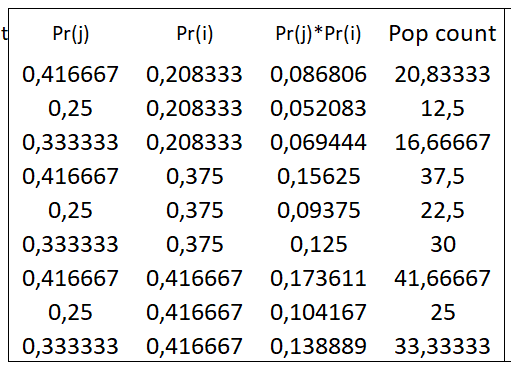
And similarly

And is given as

Given the marginal totals:



The population count method gives us the following result:



We can also use IPF to distribute the trips. Assuming a model with independence among origins and destinations means that the initial solution matrix is made up of 1s in all cells. Besides, we know the marginal totals because they were given as initial data. Taking both into account, we can run IPF procedure to obtain Tij.

First iterations, where we adjust for marginal totals of the origins is calculated as:

The next iteration, where we adjust for marginal totals of the destinations is calculated as:

We see from the table below that the IPF converges after two iterations, because there is no dependence between origins and destinations to adjust for. It also converges at the same solution as the population count method.

Table

Description automatically generated

About its utility, we consider that it is useful in order to get an initial solution, but it is not accurate, since usually there is not independence among origins and destinations.

## Section 2: Distribution new model

**Now consider a model based on the initial solution, i.e. 𝑇𝑖j=𝐴𝑖𝐵𝑗𝑂𝑖𝐷𝑗tij. Calculate the base matrix, 𝑇ijB, using this model. Do you think this is a useful model?**

In this case, the initial solution changes, since we are provided a different one, tij. However, the marginal totals remain the same because they were given as initial data. Again, taking both into account, we can run IPF procedure to obtain Tij.

Table, calendar

Description automatically generated

Figure 12. Exercise 2.2: IPF procedure

Regarding its utility, we are provided a table with initial data, so it is useful to get the initial solution. This way we have a structure, which the IPF method preserves.

## Section 3: Gravity model

**Now consider a gravity model, i.e. 𝑇𝑖j=𝐴𝑖𝐵𝑗𝑂𝑖𝐷𝑗f(gcij). A first question is what cost function to use. To find a suitable cost function, you have additional data on travel to/from zones 1, 2, and 3. These data are in the file: Distribution\_data\_pf2.xlsx. One way to estimate a cost function is to regress ln(𝑇) on generalized cost, 𝑔c. You are welcome to use other approaches. Following this you should calculate the base matrix, 𝑇ijB, using the gravity model. Do you think this is a useful model?**

In this case, the initial solution changes again, whereas margin totals stay the same. The gravity model is similar to the IPF model, only here we have a cost function.

and are balancing factors that ensure that the constraints

and

Are satisfied. It can be shown that

and

The cost function can be of multiple forms. We have a number of generalized cost functions that have been shown to be valid. We will assume that in our case, the cost function is an exponential function

If we then take the logarithm of both sides of our model (equation 1), we get the following expression

Given the data-set that we have, we can regress the cost terms on the natural logarithm of our number of trips.

To verify that the cost-function we have chosen is accurately describing the data, we can examine the residuals

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

And influential outliers:

Chart, scatter chart

Description automatically generated Chart

Description automatically generated

It seems like the regression has reasonably approximated the variance in the data.

Results for k and γ parameters were:

Chart, waterfall chart

Description automatically generated

Figure 13. Exercise 2.3: k and gamma values

We can now calculate an initial solution, from the given cost-datapoints.

Then we use the IPF fitting algorithm, where we adjust for marginal totals

Table

Description automatically generated

Figure 14. Exercise 2.3: IPF procedure

## Section 4: Evaluate models and select the best one

**Assume one of your colleagues has found the correct trip matrix for the base year.**

Diagrama, Esquemático

Descripción generada automáticamenteTabla

Descripción generada automáticamente

Figure 15. RMSE formula Figure 16. Correct trip matrix for the base year

**Try to evaluate the methods you applied in exercises 1-3 against the correct trip matrix. One way to compare them could be using the root mean square error (RMSE). Argue which of the three methods, you prefer considering that you in question 5 will be asked to investigate what happens in Labtown when a cross-city tunnel is opened between zones 2 and 3.**

Tabla

Descripción generada automáticamente con confianza media

Figure 17. Exercise 2.4: Comparison of models by RMSE

Figure 18. Exercise 2.4: Graphical comparison of models

By both RMSE (Figure X) and the graphical comparisons (Figure X) can be said that model 2 is the one which is closer to the *correct solution* provided in Figure X.

## Section 5: Prediction on best model

**Use your preferred model from 1-3 to predict the number of trips in the future scenario where a cross-city tunnel is opened between zones 2 and 3 leading to lower generalised cost between these two zones. Note that some of the other generalised cost values have been increased due to rising congestion in the scenario. Comment on your results.**

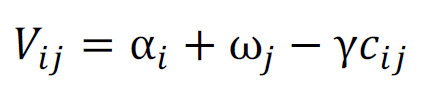
The IPF model retains the correlation between origins and destinations from a given survey. In a case like the one described, where the infrastructure connections change, this is not always a desirable outcome. Fortunately, we are given information on new marginal totals and a generalized cost structure.

Tabla

Descripción generada automáticamenteTabla

Descripción generada automáticamente con confianza media

We may isolate Ai and Bj in the gravity model, by reformulating it into a utility function

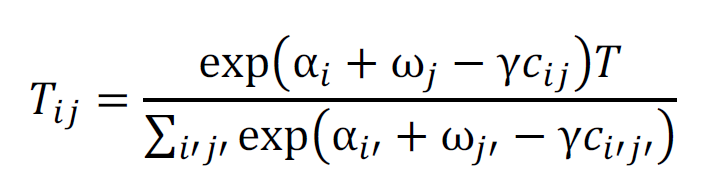


Where is utility related to the origins and is utility related to the destination. And we get the probabilities as

Diagram

Description automatically generated

And if we multiply with trips, we get number of trips

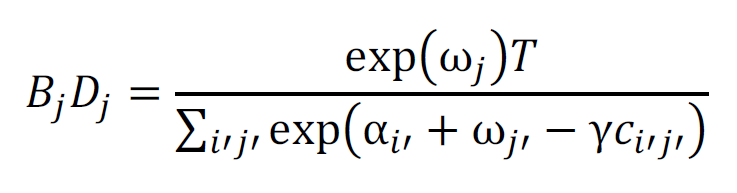


And AiOi and BjDj can be isolated as

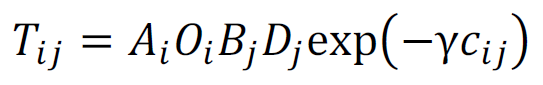
A picture containing logo

Description automatically generated

And



Then we may represent the model as



It is not possible to state Ai and Bj explicity, as it is not possible to state and explicitly. They can however be approximated by iterative calibration algorithms.

In our case, we may just apply the generalized cost-function to the cost-values, use this is an initial solution and use IPF to iteratively fit the solution to the marginal totals

