Coding assignment (QSM class)*

Hyoungchul Kim†
The Wharton School, University of Pennsylvania

Link to Latest version

Last updated September 4, 2025

Abstract

This is a coding assignment for the QSM class.

Keywords: 3 to 6 keywords

*Some footnotes on the title...

†The creator. We want to thank...

1 Use of programming language

For questions 1-6, I used `R` programming language. I also used `R` for most of the data cleaning and manipulation. For other questions that requires more computational intensive analyses (modeling, etc.), I used `julia` programming language.

2 Main data information

I am using the 2022 LODES data from LEHD for my analysis. I obtained the raw data for Philadelphia county from the LEHD website and aggregated the data into tract-tract level. The process of cleaning the raw data was conducted using the source code: `src/R/01_clean_raw_data.R`.

Thus my primary analysis zone will be bilateral commuting flow within Philadelphia county for the year 2022. This means I will not be considering the flow where either the origin or the destination is outside of Philadelphia county.

Q1

As I mentioned in Section 2, I obtained bilateral commuting flow data from LEHD LODES for the year 2022. I also downloaded supporting data on the locations of the tracts or blocks underlying the the data. I downloaded them in the `input` folder. You can also use `make raw` command to automatically download the raw data.[1]

Q2

For distance, I will use the distance between the centroids of the origin and destination tracts. For this, I used `sf` package and `tigris` package in R to calculate the distance between the centroids of the origin and destination tracts. The source code is available in `src/R/02_calculate_distance.R`. I also retained the fixed effects estimates in the Appendix.

Q3

I estimated the following linear model:

$$\log(N_{ij}) = \theta_i + \lambda_j + \kappa d_{ij} + \varepsilon_{ij}$$

The estimation results are reported in the following table:

[1] Note that if you want to re-download the raw data, you must first use `make clean` command to remove the raw data in the `input` folder. Also note that if the original data was updated during the course of the semester, there is a likelihood that the new downloaded data might have different data structure that could affect the analysis.

Table 1: Estimation results

	Log of commuting flow
distance_km	-0.07^{***}
	(0.00)
Num. obs.	73326
Num. groups: w_tract	406
Num. groups: h_tract	408
R^2 (full model)	0.61
R^2 (proj model)	0.19

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Note: This table presents estimates results of the equation 2. `distance_km` is the estimates of the semi-elasticity of the travel time (or distance). The values in the paranthesis in the table is the standard error.

You can see that the estimates of the semi-elasticity of the travel time (or distance) is sensible as it is negative (-0.07). This indicates that the travel time (or distance) has a negative correlation with the commuting flow.

Q4

After including all the ij pairs and adding in zero commuting flows, I estimated the PPML model as follows:

$$\log(\mathbb{E}[N_{ij}]) = \theta_i + \lambda_j + \kappa d_{ij}$$

The estimation results are reported in the following table (I also retained the fixed effects estimates in the Appendix):

Table 2: Estimation results of the PPML model

	PPML
distance_km	-0.12^{***}
	(0.01)
Num. obs.	165648
Num. groups: w_tract	406
Num. groups: h_tract	408
Pseudo R^2	0.64

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Note: This table presents estimates results of the equation 2. `distance_km` is the estimates of the semi-elasticity of the travel time (or distance). The values in the paranthesis in the table is the standard error.

We can clearly see that the etimates of the semi-elasticity of the distance differs from the

estimates in the previous question. The absolute size of the estimates is larger in the PPML model. Intuitively, I think this occurs because the PPML model incorporates the zero commuting flows that were neglected in the linear model. Since the zero commuting flows were not fully incorporated in the linear model, the estimates in the linear model were underestimating the effect of the travel-time cost on the commuting flow.

Q5

1. For ii pairs, we could add minimum distance from the centroid to the edge of the polygon geometry as the distance measure. While this is not a perfect measure, it still should work as a proxy for measuring the relative size of certain distance from traveling within the same tract. This also solves the zero distance issue. The only problem might be that this method could be bit ad-hoc and not consistent as we are just additional additional values only onto the ii pairs and not consider the ij pairs.

Using this method, I get the following estimates for linear and PPML models:

2. Another way would be to change the definition of the distance measure so that ii pairs will not have zero values. For example, we re-define the distance between two polygons A and B as follows: distance from the centroid of A to the centroid of B plus the minimum distance from the centroid of A to the edge of the polygon A plut the minimum distance from the centroid of B to the edge of the polygon B. With this new definition, ii pairs will no longer have zero values since we would be using diameter of the polygon as the distance measure.

Using this method, I get the following estimates for linear and PPML models:

References

APPENDIX

Appendix A