Final Practical

Kyle Lago

HLTH 661

20658108

06-Aug-2023

1. A map of the city of toronto

   Description automatically generated
2. While analyzing global clustering with Moran’s I method, we found that there is a moderate to strong positive spatial autocorrelation (Moran’s I: 0.656) within this dataset.

A screen shot of a graph

Description automatically generated

1. A screenshot of a computer screen

   Description automatically generated

As we can see from the screenshots of our LISA2 cluster and LISA significance map, there are significant (p-value <0.05 or lower) clusters of COVID-19 case rates per 100K residents within our Toronto dataset. We can see that the northwest section of the map contains significant High-High clusters, which indicate higher COVID-19 rates than other regions.

Alternatively, we see the presence of Low-Low clusters in the middle of the map, with one large region in the southeast, which indicates lower COVID-19 rates than other regions.

1. In order to utilize Geoda to perform a spatial lag regression to explore associations between a dependent variable and one independent variable, we would need the following inputs and outputs:
   1. Dependent variable (Input)
      1. In our analysis we are choosing the COVID-19 Case Rates per 100K residents (Ratep\_ople). This is our health indicator that we are interested in examining if one independent variable (predictor variable) has an associated relationship with it. That is, if our predictor variable increases or decrease is there a corresponding increase or decrease in our dependent variable. If there is, it may be that the two are related significantly.
   2. Independent variable (input)
      1. In our analysis we have four independent variables (insta\_HsTO, depri\_HsTO, depen\_HsTO, ethni\_HsTO) that we are interested in analyzing. Independent variable inputs are required for regression analysis. Without them there are no relationships that we can explore with the dependent variable. These are explanatory variables that could potentially explain the increase or decrease in COVID-19 case rates per 100k residents in our dataset. These predictor variable inputs in our dataset are related to marginalization measures indicated by Public Health Ontario.
   3. Spatial Weights Matrix (input)
      1. The spatial weights matrix is required for running a spatial lag regression. This matrix defines the spatial relationships between regions and helps identify spatial autocorrelation or dependency in our data. Without the spatial weight’s matrix, we cannot run an effective spatial lag regression.
   4. Spatial Lag Regression Results (output)
      1. After we indicate the appropriate inputs, we can “run” our spatial lag regression. In GeoDa, this will produce a text file that displays the results of our spatial lag regression. This will include relevant statistics that will allow us to determine if our predictor variable is associated with our dependent variable, and if there is any spatial autocorrelation or dependencies involved with this association. In our model, they are as followed:

* R-Squared – This tells us how much of the total variance is attributed to the predictor variables in our model. It can also be understood as how much our predictor variables can explain what is happening with our dependent variable.
* Log likelihood and AIC criterion – These help assess how good our model is relative to other models. If we were to run a similar regression with alternative methods (i.e. spatial error regression), we could use these to decide which model is best.
* Regression Coefficients – these help us quantify how much a one unit change in our predictor variable affects our dependent variable.
* P-Value – This helps us determine if findings in our model are statistically significant. If we produce p-values less than our significance levels thresholds, we can conclude that the finding for that statistic is unlikely to be due to random chance.
* Bruesch-Pagan Test – this helps determine if we have heteroskedasticity present in our model. If heteroskedasticity is present, this would be reflected in a significant value for the test. This means that there are non-random patterns among our residuals.
* Likelihood Ratio Test (LRT) – this test helps us determine whether spatial autocorrelation or spatial dependencies are present in our model. A value that is lower than our significance threshold would indicate that there is spatial autocorrelation in this model.
  + This can also be seen in the coefficient of our spatial lag term that we’ve included as part of the model.
* Model Predicted Values – GeoDa produces a table beneath the model summary that shows the predicted value, residual value and predicted error for each observation (i.e. region/area) in our model. These can be helpful if we are using our model to predict data in regions that may not have data collected.

1. Before we can begin to run our spatial lag regressions, we need to observe if our data is normally distributed. Our health indicator variable, “Ratep\_ople”, must be normally distributed for us to trust the results of our OLS spatial lag regression. We can do this by viewing a histogram of the data.

A graph of different colored bars

Description automatically generated with medium confidence

As we can see, we have the right skew in our data, and it is not normally distributed. One way we can try to obtain normality is to transform our dependent variable logarithmically. By taking the log of “Ratep\_ople”, we can see that there is distribution of the data that is closer to a normal distribution.

A graph of different colored bars

Description automatically generated with medium confidence

Now we can begin running our spatial lag regressions.

* 1. Insta\_HsTO Model

A screenshot of a computer

Description automatically generated

According to our spatial lag regression model, the insta\_HsTO explains approximately 61% of the variance in COVID-19 Case Rates per 100K. In this model, A one-unit increase in our spatially lagged dependent variable “W\_LOG\_RateCV” is associated with a 0.832123 increase in the log-transformed Case Rates of COVID-19 per 100K residents. With a significant spatial weight term, as well as significant likelihood ratio test, this means that there is spatial autocorrelation present in our model. The case rates in some areas can have a significant effect on the case rates in other areas.

However, in our model, the residential instability score was not significant. This means that in our model, there is no strong evidence of a significant relationship between residential instability score and COVID-19 case rates. We should also disregard the high R-Squared in this model considering our sole predictor variable in insignificant.

* 1. Depri-HsTO Model

A screenshot of a computer

Description automatically generated

According to our spatial lag regression model, the depri\_HsTO explains approximately 71% of the variance in COVID-19 Case Rates per 100K. In this model, A one-unit increase in our spatially lagged dependent variable “W\_LOG\_RateCV” is associated with a 0.520568 increase in the log-transformed Case Rates of COVID-19 per 100K residents. With a significant spatial weight term, as well as significant likelihood ratio test, this means that there is spatial autocorrelation present in our model. The case rates in some areas can have a significant effect on the case rates in other areas.

A one-unit increase in material deprivation score is associated with a 0.29674 increase in the log-transformed COVID-19 case rates per 100K residents. In the original scale, a one-unit increase in material deprivation score is associated with approximately 34.5% in the Case Rates of COVID-19 per 100K residents. The remaining portion of the variance can be explained by the case rates in neighbouring areas.

Our Bruesch-Pagan test result was insignificant, indicating that there is no heteroskedasticity in our data. This allows us to be more confident in the regression coefficient and confidence interval estimates produced by our model.

* 1. Depen\_HsTO Model

A screenshot of a computer

Description automatically generated

According to our spatial lag regression model, the depen\_HsTO explains approximately 59.2% of the variance in COVID-19 Case Rates per 100K. In this model, A one-unit increase in our spatially lagged dependent variable “W\_LOG\_RateCV” is associated with a 0.797881 increase in the log-transformed Case Rates of COVID-19 per 100K residents. With a significant spatial weight term, as well as significant likelihood ratio test, this means that there is spatial autocorrelation present in our model. The case rates in some areas can have a significant effect on the case rates in other areas.

However, in our model, the dependency score was not significant. This means that in our model, there is no strong evidence of a significant relationship between dependency score and COVID-19 case rates. We should also disregard the high R-Squared in this model considering our sole predictor variable is insignificant.

* 1. Ethnic\_HsTO Model

A screenshot of a computer

Description automatically generated

According to our spatial lag regression model, the ethni\_HsTO explains approximately 68.8% of the variance in COVID-19 Case Rates per 100K. In this model, A one-unit increase in our spatially lagged dependent variable “W\_LOG\_RateCV” is associated with a 0.669464 increase in the log-transformed Case Rates of COVID-19 per 100K residents. With a significant spatial weight term, as well as significant likelihood ratio test, this means that there is spatial autocorrelation present in our model. The case rates in some areas can have a significant effect on the case rates in other areas.

A one-unit increase in ethnic concentration score is associated with a 0.243341 increase in the log-transformed COVID-19 case rates per 100K residents. In the original scale, a one-unit increase in material ethnic concentration score is associated with approximately 27.6% in the Case Rates of COVID-19 per 100K residents. The remaining portion of the variance can be explained by the case rates in neighbouring areas.

Our Bruesch-Pagan test result were significant, indicating that there is heteroskedasticity present in our model. While this does not take away from the explanatory power of the model, we trust our regression coefficients and confidence intervals less in their predictive capabilities. To improve this model, heteroskedasticity should be addressed.

* 1. Of the four domains that could predict the health indicator, ethnic concentration and material deprivation scores were best. We can make this determination by reviewing the regression results, and seeing these were both significant predictors of COVID-19 Case Rates.

However, of these two predictors the most precise would be the material deprivation score as there was no heteroskedasticity present in the model. To trust the predictions from the ethnic concentration model, we would need to employ methods to address the heteroskedasticity that is present. These methods could affect other values in our model, so it would not be accurate to rely on them currently. For the scope of this assignment, I would choose the material deprivation score as the most accurate predictor of COVID-19 case rates in the City of Toronto.

1. I would recommend these public health intervention strategies:

* Target interventions in neighborhoods in the northwest area of the city.
* Target interventions in these neighborhoods with high ethnic concentration.
* Target the poorest neighborhoods in this area of the city.

Therefore, implementing public health measures in the northwest area of the city would be the most beneficial to help combat COVID-19. These are neighborhoods with high levels of ethnic residents with low socioeconomic status. They are disadvantaged in accessing services and education to prevent the spread of COVID-19, which is likely why their cases are so high.