Impact of Socio-Economic Factors on Travel Times to Auraria Campus in Denver

Ryan H. Peterson

Department of Mathematics and Statistics
Metropolitan State University of Denver

Spring 2021

Outline

- Motivating Factors and Background Information
- Data Collection and Preparation
- Statistical Methods and Model Building
- Validation and Results
- Conclusion and Future Goals

Motivating Questions

- Are there socio-economic factors that influence an individual's or household's decision to live in certain places, with respect to a central place?
- Do socio-economic factors that affect a student's ability to access Auraria Campus?
- If there is a correlation between socio-economic factors and travel times, does it encompass the entire Metropolitan Statistical Area or is there some other boundary?

Further Motivations

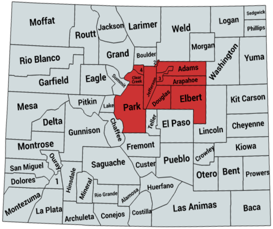
- There is a high degree of economic and social integration across urban spaces.
- I hypothesize that socio-economic integration across an urban space can be used to better define Metropolitan Statistical Areas.
- This is an exploratory observational study.

Infrastructure Networks

- Urban spaces are navigated via infrastructure networks.
- Therefore, socio-economic interactions occur on or near these infrastructure networks.
- When determining travel times to a point in an urban space needs to be calculated by navigating infrastructure networks.

Source: OpenStreetMaps

Metropolitan Statistical Areas



Source: MapChart

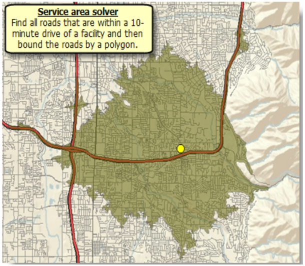
- Metropolitan Statistical Areas (MSA) are comprised of central and outlying counties.
  - Outlying counties are considered to have a high degree of social or economic integration based on commuting into the central counties.
- Socio-economic data was collected from the Denver-Aurora-Lakewood MSA.

Data Collection and Selection

- Socio-economic data from American Community Survey 2015.
  - Split into Economic, Housing, Demographic, and Social Data.
  - Organized by Census Tract and County codes.
- Select counties that are in the Denver MSA.
  - Create reference ID codes that include County and Census Tract.

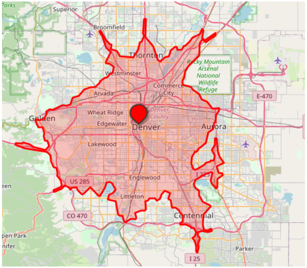
Travel Time Data

- Travel time is based on private vehicle use and the navigation of road (infrastructure) networks.
- To determine travel times we needed to perform a network analysis of the Denver MSA.
  - This is typically done using GIS software (either ArcGIS or QGIS).
  - Alternatively, travel times can be determined through Google Maps Platform.



Source: ArcGIS Service Area Analysis

Open Route Service



Source: OpenRouteService Isochrone Map

- Open source software.
- Provided a python library that include Time-Distance Matrix.
  - The Denver MSA has approximately 620 census tracts.
- Navigated from the geometric center of population for each Census Tract to Auraria Campus.

Variable Selection

- Selected several likely variables from each ACS data set (Economic, Housing, Demographic, and Social).
- Initial variable reduction done through collinearity comparisons in R.
  - Using the R pairs function we could visually identify collinearities.
  - Then by examining a Pearson Correlation Matrix, we were able to numerically eliminate an additional collienarity.

Selected Variables

From the work of variable selection and reduction, the following socio-economic factors were chosen as predictors:

X_1 is the Labor Participation rate of a given census tract.
X_2 is the Percent of households with one or more person under 18.
X_3 is the Percent of population with income below the poverty line.
X_4 is the Percent of population that do not identify as only White.
X_5 is the Percent of housing structures that are single family homes.
X_6 is the Percent of the population has the educational attainment of a high school diploma.

Model Building

- Since all predictors and response variables are numeric, a general linear model was selected.
- Based on some intuition and hypothesized interaction, several interactions were identified.
- Therefore, the model should be in the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_8 X_3 X_4 + \epsilon$$

Initial Model

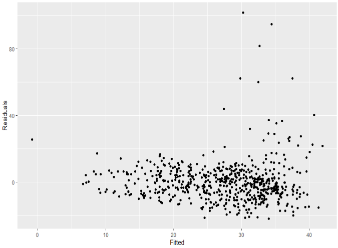
Summary Table

```
Residuals:
    Min      1Q  Median      3Q     Max
-25.303  -6.731  -1.462   4.627 100.732

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     50.336306   9.086481   5.540 4.53e-08 ***
LABOR_PARTICIPATION             -0.371592   0.128990  -2.881  0.00411 **
HOUSE_UNDER18                   -0.757363   0.270031  -2.805  0.00520 **
BELOW_POVERTY                   -0.921203   0.118873  -7.749 3.93e-14 ***
POP_NONWHITE                    -0.338001   0.070862  -4.770 2.31e-06 ***
SINGLE_FAMILY_HOMES              0.075006   0.025606   2.929  0.00353 **
HS_EDUCATION                     0.311127   0.067034   4.641 4.25e-06 ***
LABOR_PARTICIPATION:HOUSE_UNDER18 0.012379   0.003936   3.145  0.00174 **
BELOW_POVERTY:POP_NONWHITE       0.018451   0.003472   5.315 1.51e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.33 on 603 degrees of freedom
Multiple R-squared:  0.2746,    Adjusted R-squared:  0.265
F-statistic: 28.53 on 8 and 603 DF,  p-value: < 2.2e-16
```

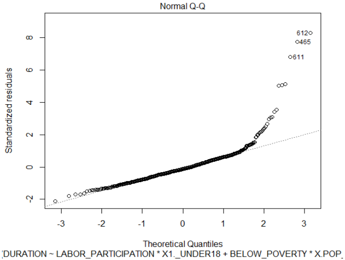
Figure: 1.1

Model Problems

- However, there is a problem with this model.
  - There is not constant variance.
  - The residuals do not appear to be normally distributed.



Residuals vs. Fitted



Q-Q Plot

Response Transformations

- Due to lack of normality in our error we needed to find a transformation.
- I decided to use the Box-Cox Transformation.

Where the transformation is defined as:

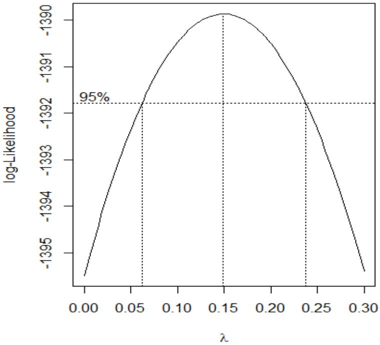
$$g(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

And the maximum likelihood estimator of λ is:

$$L(\lambda) = -\frac{n}{2} \log(\hat{\sigma}^2(\lambda)) + (\lambda - 1) \sum \log y_i$$

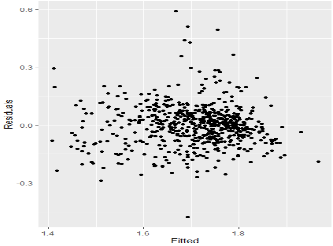
Transformation

- From the Box-Cox Transformation $\lambda \approx 0.15$.
- Therefore, we can use a transformation of $g(y) = y^{1/6}$, since $\lambda = \frac{1}{6}$ is within the 95% confidence interval.

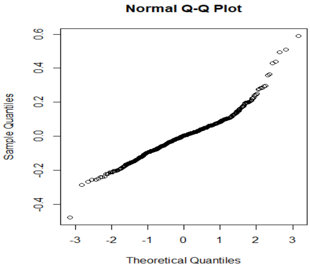


Log-Likelihood of λ.

New Model Plots



Transformed Residuals vs. Fitted



Transformed Q-Q Plot

New Linear Model

The new linear model uses the full transformation of y, where $g(y) = \frac{y^{1/6}-1}{1/6}$,

- This transformation yields an increased R^2 value of 0.422.
- However, there is not longer significance for the predictor X_5 (percent single family homes).

New Model

Summary Table

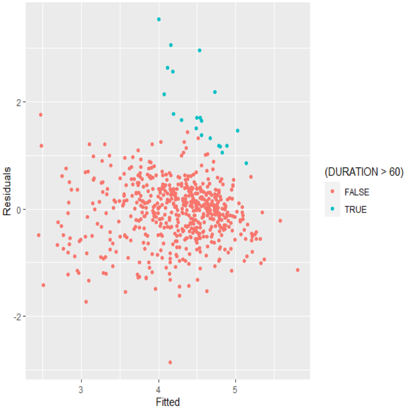
```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.8656 -0.3923  0.0201  0.3420  3.5370 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       5.9963804  0.4941588  12.135  < 2e-16 ***
LABOR_PARTICIPATION              -0.0292177  0.0070150  -4.165 3.57e-05 ***
HOUSE_UNDER18                    -0.0347724  0.0146854  -2.368 0.018207 *  
BELOW_POVERTY                    -0.0758711  0.0064648 -11.736  < 2e-16 ***
POP_NONWHITE                     -0.0187313  0.0038537  -4.861 1.49e-06 ***
SINGLE_FAMILY_HOMES               0.0001134  0.0013926   0.081 0.935141    
HS_EDUCATION                      0.0233888  0.0036456   6.416 2.84e-10 ***
LABOR_PARTICIPATION:HOUSE_UNDER18 0.0007811  0.0002140   3.649 0.000286 ***
BELOW_POVERTY:POP_NONWHITE        0.0012006  0.0001888   6.359 4.02e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6705 on 603 degrees of freedom
Multiple R-squared:  0.4223,    Adjusted R-squared:  0.4147 
F-statistic: 55.11 on 8 and 603 DF,  p-value: < 2.2e-16
```

Figure: 1.2

Validation

- For validation we set up training and testing data.
- Testing data was between 10% and 20% of our sample size.
- The Mean Absolute Error (MAE) was around 0.5.
  - Which can be transformed back into our original unit of Travel Time (minutes).
  - MAE $\approx$ 1.5 minutes.

A Better Model

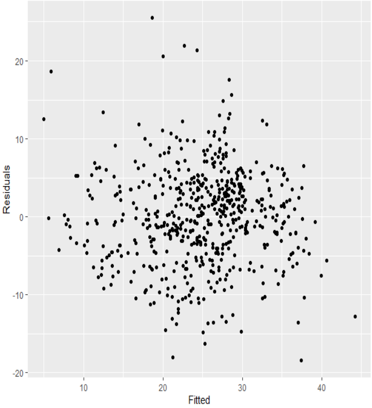


Residual vs. Fitted, with travel times $>$ 60 minutes highlighted.

- However, we can do better.
- Census tracts with very long travel times may indicate an absence of socio-economic integration.
- By "shrinking" the boundary of our model, with a maximum travel time duration, we may be able to improve upon our earlier model.

Maximum Duration

- With a maximum travel time duration of approximately 46 minutes, we don't need a transformation.
- The Residual vs. Fitted plot improves.
- R^2 goes up to 0.54.



Residual vs. Fitted for travel times $<$ 46 minutes.

Duration Boundary Model

Summary Table

```
Residuals:
     Min       1Q   Median       3Q      Max 
-18.4484  -3.7696   0.1243   3.7828  25.4815 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     46.727660   4.626036  10.101  < 2e-16 ***
LABOR_PARTICIPATION             -0.343758   0.065699  -5.232 2.36e-07 ***
HOUSE_UNDER18                   -0.214231   0.137649  -1.556   0.1202    
BELOW_POVERTY                   -0.804175   0.061017 -13.180  < 2e-16 ***
POP_NONWHITE                    -0.095693   0.036509  -2.621   0.0090 ** 
SINGLE_FAMILY_HOMES             -0.070254   0.013460  -5.220 2.52e-07 ***
HS_EDUCATION                     0.072475   0.035163   2.061   0.0398 *  
LABOR_PARTICIPATION:HOUSE_UNDER18  0.008763   0.002003   4.375 1.44e-05 ***
BELOW_POVERTY:POP_NONWHITE       0.009492   0.001781   5.329 1.43e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.19 on 566 degrees of freedom
Multiple R-squared:  0.5416,	Adjusted R-squared:  0.5351 
F-statistic:  83.6 on 8 and 566 DF,  p-value: < 2.2e-16
```

Figure: 1.3

Conclusions

- There appears to be a connection between certain socio-economic factors and the time it takes to travel to Auraria campus.
- By creating a duration boundary we eliminate a the need for a transformation and create a model that better reflects the socio-economic integration of an urban area.

Future Goals

- Investigate the sensitivity of the model.
- Do the same process with non-private vehicle transportation (public transport, walking, biking, etc.) to see if the model still holds.
- I would like to apply this model to other MSAs to see if it works on other urban areas.
  - See if there is a consistent or scaling travel time duration boundary across urban areas.

Thank you!

Any Questions or Comments?