Stat 333 - Applied Linear Regression Analysis Project:

Predicting the Natality with Means of Transportation

Tess Steplyk

University of Wisconsin-Madison

1. Introduction

In the past five years, more laws and bills have been written to restrict women's healthcare than the previous 15 years combined. The more than 200 abortion restrictions enacted since the beginning of 2011 and federal funding cuts to low income women's healthcare centers have caused for mass closures of these centers. With fewer women's healthcare centers spread across the United States, women may now be forced to drive 200 miles roundtrip for her basic healthcare needs.

If a woman really wants her healthcare she would take that drive for things like affordable mammograms and sexually transmitted disease testing to contraceptives and abortion, but not everyone has the ability to take such a long trip. The women this directly effects are low-income women who cannot afford traditional gynecologists or do not have health insurance. So if these women are unable to afford these two things, can she afford a car to take her to one of these centers? If the woman takes a car as her means of transportation to work, and therefore would have access to a car to drive to a women's healthcare center, will show she will give birth to less children.

2. Methods

2.1 Data Sources

A dataset was retrieved Center for Disease Control and Prevention for the project. The data collection was performed by the National Center for Health Statistics (NCHS) from January 2010 to December 2014. The CDC dataset records the number of births, one for each individual living baby, occurring in the United States to residents and non-residents. This is the natality, or birth rate, dataset with each used that gives the total number of babies born in each United States county over the population of 100,000 persons or more, the counties that have left are listed as “Unidentified Counties” and have been removed. There is a total of 19,826,074 babies and all the data has been derived from birth certificates from 2010 to 2014. [http://wonder.cdc.gov/natality-current.html]

Means of transportation data was collected by the American Community Survey (ACS). by the United States Census Bureau which was an ACS 5-year estimate of workers of driving age, age 16 years or older. The estimate is from 2010 to 2014, the same years as my natality dataset. The data was taken from a questionnaire in 2013, and therefore is not an exact number. After cleaning the data, the variables that were most useful in my project became drive (2), public (3), bike (6), and walk (7).
[https://www.socialexplorer.com/data/ACS2009_5yr/metadata/?ds=ACS09_5yr&table=B08301]

	NAME	DESCRIPTION
1	**total**	**total workers**
2	**drive**	**drive car, carpool**
3	**public**	**public transportation**
4	**taxi**	**taxi cab**
5	**motor**	**motorcycle**
6	**bike**	**bicycle**
7	**walk**	**walk**
8	**other**	**other**
9	**home**	**work from home**

2.2 Data Cleaning

Prior to trying to fit any model I decided to remove the variables that did not fit with the end goal of proving women who have access to cars (*drive*) are going to have a lower natality rate. Therefore, I decided to remove working from *home* because that means of transportation to work does not cross over to how that woman would get around regularly. *Taxi* and *motorcycle* were focused on too specific of counties. *Other* doesn’t have a place in helping the natality goal. *Total* just does not relate to the end analysis goal.

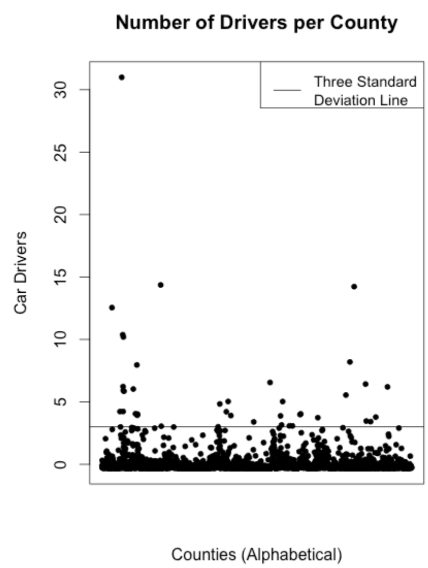
The natality dataset came in a .csv file. I removed all the unidentified counties and NA data in R.

```
> summary(ustrans)
     total              drive              public              taxi            motorcycle             bike              walk              other               home       
 Min.   :     46   Min.   :     17   Min.   :     0   Min.   :    0.0   Min.   :    0.00   Min.   :    0   Min.   :     0   Min.   :    0.0   Min.   :     0  
 1st Qu.:   4551   1st Qu.:   4039   1st Qu.:     8   1st Qu.:    0.0   1st Qu.:    0.00   1st Qu.:    0   1st Qu.:   118   1st Qu.:   31.0   1st Qu.:   187  
 Median :  10510   Median :   9530   Median :    38   Median :    0.0   Median :   18.00   Median :   18   Median :   268   Median :   89.0   Median :   429  
 Mean   :  43645   Mean   :  37653   Mean   :  2157   Mean   :   50.7   Mean   :   96.77   Mean   :  232   Mean   :  1239   Mean   :  380.1   Mean   :  1836  
 3rd Qu.:  28482   3rd Qu.:  26004   3rd Qu.:   159   3rd Qu.:   13.0   3rd Qu.:   65.00   3rd Qu.:   86   3rd Qu.:   679   3rd Qu.:  242.0   3rd Qu.:  1134  
 Max.   :4382882   Max.   :3649880   Max.   :649563   Max.   :26311.0   Max.   :11222.00   Max.   :33960   Max.   :177293   Max.   :41706.0   Max.   :204960  
```

```
> summary(nat)
                 county        birth       
 Ada County, ID      :  1   Min.   :   789  
 Adams County, CO    :  1   1st Qu.:  1967  
 Aiken County, SC    :  1   Median :  3307  
 Alachua County, FL  :  1   Mean   :  6907  
 Alamance County, NC:  1   3rd Qu.:  7198  
 Alameda County, CA  :  1   Max.   :133252  
 (Other)            :518                 
```

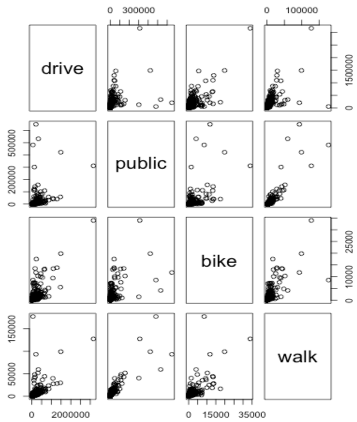
2.3 Data Choosing

When plotting *drive*, the response variable, against counties there a few outliers. Shown in the graph with random points far above the line. Since it can be assumed the outliers are particularly for larger counties and/or counties with a more condensed population I wanted to take a look at the data with and without them to see if there is a drastic change.



2.3.1 Outliers

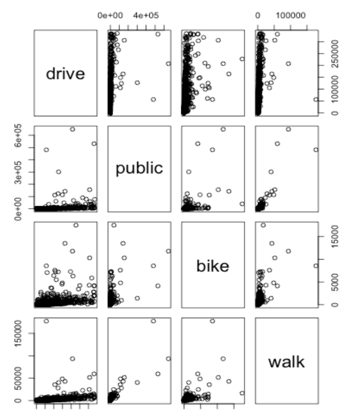
tran remains my main data frame with my response variable. Here is a summary and plot of *tran* with outliers in the data set. They can easily be seen as particularly the maximums.




```
> summary(tran)
     drive             public            bike              walk        
 Min.   :     17   Min.   :     0   Min.   :    0   Min.   :     0  
 1st Qu.:   3514   1st Qu.:     8   1st Qu.:    0   1st Qu.:   118  
 Median :   8307   Median :    38   Median :   18   Median :   268  
 Mean   :  33216   Mean   :  2157   Mean   :  232   Mean   :  1239  
 3rd Qu.:  22926   3rd Qu.:   159   3rd Qu.:   86   3rd Qu.:   679  
 Max.   :3164442   Max.   :649563   Max.   :33960   Max.   :177293  
```

tranR after my outlier removal process on *tran*. The plots are not more correlated than they were with *tran.* I chose to stay with my original data, with the outliers, because removing them did not change too much.

```
> summary(tranR)
     drive            public            bike               walk        
 Min.   :    17   Min.   :     0   Min.   :    0.0   Min.   :     0  
 1st Qu.:  3462   1st Qu.:     8   1st Qu.:    0.0   1st Qu.:   117  
 Median :  8145   Median :    36   Median :   17.0   Median :   262  
 Mean   : 25208   Mean   :  1563   Mean   :  163.7   Mean   :   987  
 3rd Qu.: 21648   3rd Qu.:   146   3rd Qu.:   82.0   3rd Qu.:   655  
 Max.   :335758   Max.   :649563   Max.   :17509.0   Max.   :177293  
```



There's not a sufficient amount information to explain the reason for high residuals, so the outliers were not removed.

3. Results

3.1 Full Model

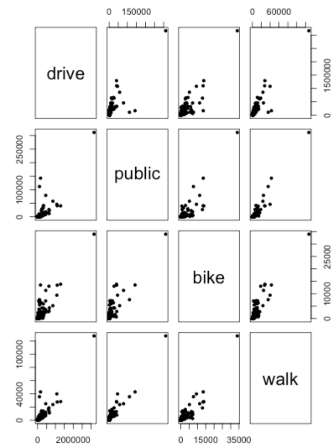
Here is a summary of statistics of the full model of my dataset with predictors for *tran*.

```
> summary(lmfit)

Call:
lm(formula = drive ~ public + bike + walk, data = tran)

Residuals:
     Min       1Q   Median       3Q      Max 
-1241699   -12611    -9560    -1951   992178 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept) 13522.1511  1129.1384   11.98 <0.0000000000000002 ***
public         -1.6143     0.1037  -15.56 <0.0000000000000002 ***
bike           50.9210     1.4700   34.64 <0.0000000000000002 ***
walk            9.1723     0.5041   18.20 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61570 on 3217 degrees of freedom
Multiple R-squared:  0.6359,	Adjusted R-squared:  0.6356 
F-statistic:  1873 on 3 and 3217 DF,  p-value: < 0.00000000000000022
> lmfit

Call:
lm(formula = drive ~ public + bike + walk, data = tran)

Coefficients:
(Intercept)       public         bike         walk  
  13522.151       -1.614       50.921        9.172  
```

3.2 Correlation matrix

	drive	public	bike	walk
drive	1.0000000	0.4382820	0.7734112	0.6568674
public	0.4382820	1.0000000	0.5848256	0.8664151
bike	0.7734112	0.5848256	1.0000000	0.7644823
walk	0.6568674	0.8664151	0.7644823	1.0000000

There’s some inter-correlations among predictors shown in the table above and the scatterplot matrix to the bottom left. There are strong relations with *public* and *walk* (r = 0.8664) and *bike* (r = 0.7644), which suggest some potential needs for multi-collinearity remediation. On the other hand, the response, drive, is correlated with many other predictors, such as *walk* (r = 0.6568 and *bike* (0.7734). The results imply a regression model might be practicable method to predict *drive***.**



3.3 Predicting

When trying to predict *drive* for plots and natality. Using *public* = -1.6143, *bike* = 50.921, and *walk* = 9.1723 to look at each variable and map.

```
> predictions = predict.lm(lmfit, newdata)
> predictions
       1
35.15282
```

```
> print(newlm)
       fit     lwr      upr
1 35.15282 33.7179 36.58773
```

3.4 Variables - Stepwise

Stepwise procedure with both forward and backward direction selected the full model results in the next section. Still prefer the chosen model model selected from all regression procedure, its implementation on adjusted R-squared is comparable with the full model even without some of the variables. Stepwise helps with choosing and double checking you are choosing the right variables. All three are good for the final result.

```
Start:  AIC=74294.78
drive ~ 1

         Df      Sum of Sq            RSS   AIC
+ bike    1 20038733203777 13461618145410 71360
+ walk    1 14454555370132 19045795979054 72478
+ public  1  6435120732657 27065230616530 73610
<none>                     33500351349187 74295

Step:  AIC=71360.15
drive ~ bike

         Df      Sum of Sq            RSS   AIC
+ walk    1   346996333256 13114621812153 71278
+ public  1    10019948536 13451598196874 71360
<none>                     13461618145410 71360
- bike    1 20038733203777 33500351349187 74295

Step:  AIC=71278.04
drive ~ bike + walk

         Df     Sum of Sq            RSS   AIC
+ public  1  918161633177 12196460178977 71046
<none>                    13114621812153 71278
- walk    1  346996333256 13461618145410 71360
- bike    1 5931174166901 19045795979054 72478
```

```
Step:  AIC=71046.25
drive ~ bike + walk + public

         Df     Sum of Sq            RSS   AIC
<none>                    12196460178977 71046
- public  1  918161633177 13114621812153 71278
- walk    1 1255138017897 13451598196874 71360
- bike    1 4549216923619 16745677102595 72065

Call:
lm(formula = drive ~ bike + walk + public, data = tran)

Coefficients:
(Intercept)         bike         walk       public
  13522.151       50.921        9.172       -1.614
```

3.5 Results

Summary of statistics of the final model. All partial T tests as well as the overall F test were highly significant. As the p-values of each variable are less than 0.05, they are all statistically significant in the multiple linear regression model of *drive*. These models are both after normalizing the units and coefficients.

```
> summary(lmfitnorm)

Call:
lm(formula = drive ~ public + bike + walk, data = transUNS)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1736  -0.1236  -0.0937  -0.0191   9.7273

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.161e-16  1.064e-02    0.00        1
public      -3.416e-01  2.195e-02  -15.56   <2e-16 ***
bike         5.890e-01  1.700e-02   34.64   <2e-16 ***
walk         5.026e-01  2.762e-02   18.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6037 on 3217 degrees of freedom
Multiple R-squared:  0.6359,    Adjusted R-squared:  0.6356
F-statistic:  1873 on 3 and 3217 DF,  p-value: < 2.2e-16
```

```
> lmfitnorm

Call:
lm(formula = drive ~ public + bike + walk, data = transUNS)

Coefficients:
(Intercept)       public         bike         walk  
 -1.161e-16   -3.416e-01    5.890e-01    5.026e-01  
```

3.6 Residuals

The bottom right purple summary plots for the best, normalized model (*lmfitnorm).* For the most part, the residual plots for all variables and all interaction terms have no specific patterns. So, I did not consider adding interaction term to the model and transforming any variables. Based on the normal probability plot, the residuals diverge from the projected residuals under normality substantially, which is shown with large deviation from the normal line at the upper right. Outlier removal was not performed because we do not have sufficient information to explain the reason for these high residuals.

Looking at outliers now that there is a fitted full model, begin with leverage scores below.

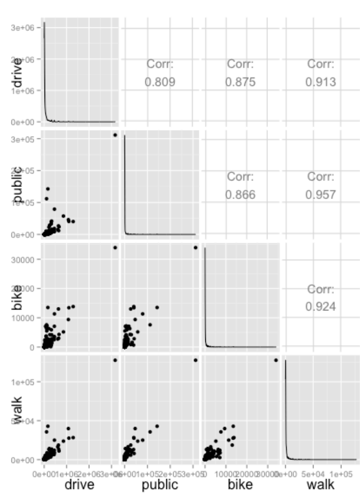


Figure 1: Normal Probability Plot

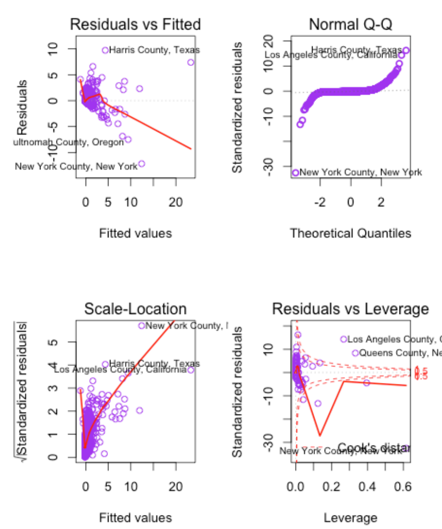
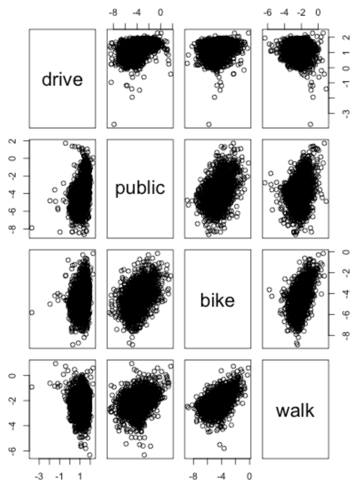
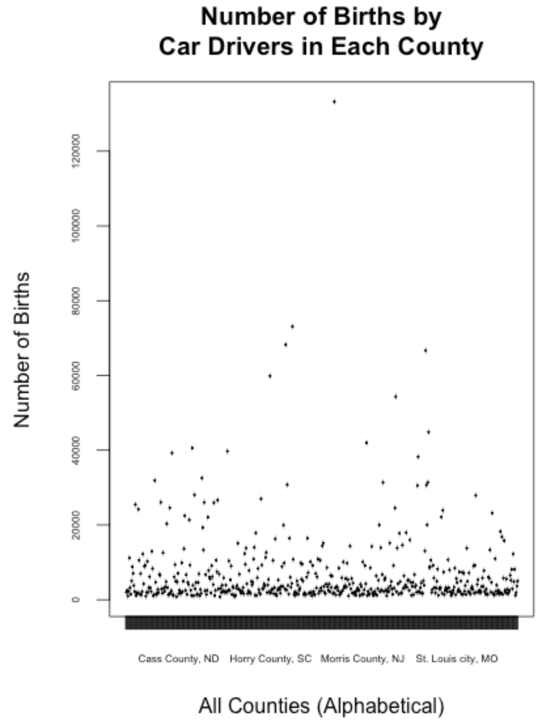


Figure 2: Leverage scores included

3.7 Natality

The natality dataset, *birth*, plotted against *drive* after *tran* was normalized by the population (shown at the bottom left). There are few outliers with *birth* and the normalized *drive* so I decided the variables do not need any obvious transforming.





4. Conclusion

Focusing particularly on x *(drive)* to find the right model, it was easy to compare with natality. Shown is how women who have access to a car can drive to a woman's health center, but the correlation is very small. There are many cofounding variables to consider with this study and I believe with adding in many different variables a better conclusion could have been drawn.

Considering the age of the mother, what kind of county she lives in, is she in New York City or the middle of Texas? How condensed the counties are also having to do with how someone gets to work. Looking specifically at someone's income and whether they can afford to buy or own a car. Confound variables such as determining if a woman could afford to take off a day of work to go to a center would be very hard to analize. Therefore, looking at just if a household owns a car would solve many of the issues my data had. Beginning with nine variables, and determining three predictors for modeling *drive* concentration in work transportation, including public, walk, and bike.

There are many limitations to my current normalized and finished model. For more in depth research, heteroscedasticity should be considered by changing the structure of the variance. I didn't want to remove any outliers because I was uncertain of why they were outliers, but with more information next time the reasons for the outliers should be considered if they were a manual error.

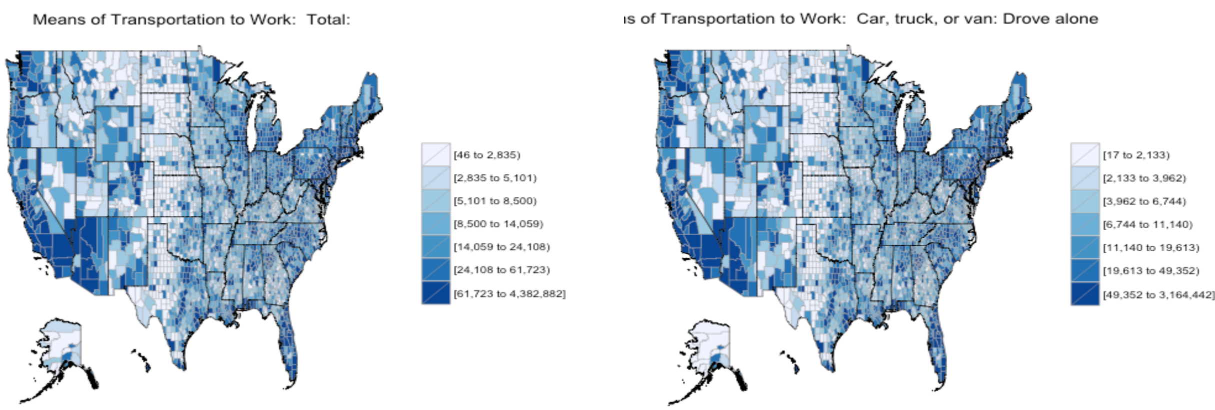


Figure 3: Maps of transportation and drive on the maps for an overview picture

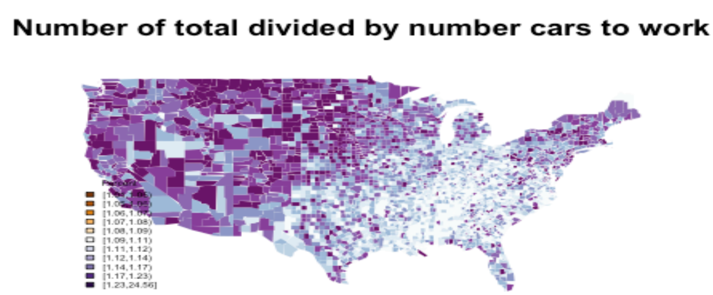


Figure 4: Ratio of workers who drive to work and who do not