Final Project Report

Multiple Regression Analysis on NHTS Phoenix-Mesa Dataset of Person and Household Trips

Miguel Cecchini do Amaral
November 21th, 2020

Florida Polytechnic University

Table of Contents

List of Tables

List of Figures

Executive Summary

The following explanatory analysis has as its guiding principle exploring the collection of facts regarding person and household trips parameters, in order to transform raw data into relevant knowledge. The dataset was derived from NHTS Phoenix-Mesa sub-sample and it contains over 297 observations for household trips and 648 observations for personal trips. The report will cover the regression analysis of the best models for both datasets and a cross-classification matrix of trip generation using techniques of statistical analysis and the statistical application software package STATA to better understand the relationship and the of the variables.

Data Dictionary

Data dictionary for person trip file:

1. **driver** : -1 appropriate skip, 1 yes a driver, 2 not a driver
2. **worker** : 1 = yes , 2 = no
3. **educ** : `-1 appropriate skip, -7 refused, 1 less than high school, 2 greater than HS
4. **hhincttl** : see household file
5. **numadlt** : see household file
6. **drvrcnt r_age :** age of person
7. **r_sex** : 1= male, 2 = female
8. **hhsize** : see household file
9. **homeown** : 1 = own, 2 = rent
10. **pertrips** : number of person trips

Data dictionary for household trip file:

1. **homeown** : 1 = own, 2 = rent
2. **hhvehcnt** : number of vehicles in household
3. **hhsize** : number of people in household
4. **drvrcnt** : number of drivers in household
5. **wrkcount** : number of workers in household
6. **numadlt** : number of adults in household
7. **trpmiles** : total number of miles traveled in household
8. **hhincttl :** -7 = Refused -8 = Don't know -9 = Not ascertained 01 = < $5,000 02 = $5,000 - $9,999 03 = $10,000 - $14,999 04 = $15,000 - $19,999 05 = $20,000 - $24,999 06 =

$25,000 - $29,999 07 = $30,000 - $34,999 08 = $35,000 - $39,999 09 = $40,000 - $44,999 10 = $45,000 - $49,999 11 = $50,000 - $54,999 12 = $55,000 - $59,999 13 = $60,000 - $64,999 14 = $65,000 - $69,999 15 = $70,000 - $74,999 16 = $75,000 - $79,999 17 = $80,000 - $99,999 18 = > = $100,000

Summary Statistics

In order to establish common ground for the starting point of data analysis, a summary statistics of all the available variables of the datasets was implemented. Table 1 through 6 and Table 9 illustrate some basal characteristics of the elements from the person trip file, such as mean, standard deviation, minimum, and maximum for the numerical data, and the frequency, percentage, and cumulative percentage for all the categorical data. The same approach was undertaken to understand the dimensions of the household trip file, as demonstrated in Table 7, 8, and 10. Those descriptive statistics are essential for the comprehension of the basic features of the data and were utilized to clarify the relevance of each variable. The results were rounded to the second decimal and there is no missing values in both datasets.

Table 1. Summary statistics hhsize, drvrcnt, r_age, numadlt, pertrips.

Variable	Obs	Mean	Std. Dev.	Min	Max
hhsize	648	3.35	1.65	1	9
drvrcnt	648	2.07	0.83	0	5
r_age	648	37.43	23.76	-8	88
numadlt	648	2.08	0.67	1	4
pertrips	648	4.59	2.39	1	16

Table 2. Tabulate statistics homeown.

homeown	Freq.	Percent	Cum.
1	525	81.02	81.02
2	123	18.98	100.00
Total	648	100.00	

Table 3. Tabulate statistics r_sex.

r_sex	Freq.	Percent	Cum.
1	320	49.38	49.38
2	328	50.62	100.00
Total	648	100.00	

Table 4. Tabulate statistics worker.

worker	Freq.	Percent	Cum.
-9	1	0.15	0.15
-1	147	22.69	22.84
1	317	48.92	71.76
2	183	28.24	100.00
Total	648	100.00	

Table 5. Tabulate statistics driver.

driver	Freq.	Percent	Cum.
-1	147	22.69	22.69
1	468	72.22	94.91
2	33	5.09	100.00
Total	648	100.00	

Table 6. Tabulate statistics hhincttl for person trip.

hhincttl	Freq.	Percent	Cum.
-8	11	1.70	1.70
-7	24	3.70	5.40
1	9	1.39	6.79
2	18	2.78	9.57
3	13	2.01	11.57
4	28	4.32	15.90
5	16	2.47	18.36
6	63	9.72	28.09
7	27	4.17	32.25
8	46	7.10	39.35
9	9	1.39	40.74
10	38	5.86	46.60
11	24	3.70	50.31
12	59	9.10	59.41
13	30	4.63	64.04
14	30	4.63	68.67
15	20	3.09	71.76
16	30	4.63	76.39
17	45	6.94	83.33
18	108	16.67	100.00
Total	648	100.00	

Table 7. Tabulate statistics hhincttl for household trip.

hhincttl	Freq.	Percent	Cum.
-8	6	2.02	2.02
-7	15	5.05	7.07
1	4	1.35	8.42
2	10	3.37	11.78
3	9	3.03	14.81
4	16	5.39	20.20
5	11	3.70	23.91
6	27	9.09	33.00
7	12	4.04	37.04
8	21	7.07	44.11
9	5	1.68	45.79
10	22	7.41	53.20
11	11	3.70	56.90
12	21	7.07	63.97
13	12	4.04	68.01
14	11	3.70	71.72
15	8	2.69	74.41
16	13	4.38	78.79
17	20	6.73	85.52
18	43	14.48	100.00
Total	297	100.00	

Table 8. Summary statistics hhvehcnt, hhsize, drvrcnt, wrkcount, defrfnumnumadlt, trpmiles.

Variable	Obs	Mean	Std. Dev.	Min	Max
hhvehcnt	297	1.89	1.10	0	7
hhsize	297	2.65	1.43	1	9
drvrcnt	297	1.85	0.79	0	5
wrkcount	297	1.26	0.99	0	5
numadlt	297	1.90	0.65	1	4
trpmiles	297	117.32	266.39	-64	3164

Table 9. Tabulate statistics educ.

educ	Freq.	Percent	Cum.
-7	1	0.15	0.15
-1	154	23.77	23.92
1	52	8.02	31.94
2	127	19.60	51.54
3	11	1.70	53.24
4	118	18.21	71.45
5	30	4.63	76.08
6	85	13.12	89.20
7	5	0.77	89.97
8	65	10.03	100.00
Total	648	100.00	

Table 10. Tabulate statistics homeown.

homeown	Freq.	Percent	Cum.
1	236	79.46	79.46
2	61	20.54	100.00
Total	297	100.00	

Data

In the first moment of the analysis, a cross-classification matrix of household trip rates by household size, number of vehicles, and the number of workers were undertaken, as shown in Figures 11, Figure 12, and Figure 13. All variables were manipulated to ensure clear results, limiting the number of value options up to 4. In figure 11, for example, the represented household size can be 1, 2, 3, and number 4 represents 4 persons in the household or more. The same logic was implemented for the *wrkcount* table (0, 1, 2, 3+) and the *hhvehcnt* table (0, 1, 2, 3, 4+).

This way, we were able to visually observe how the data behaves facing each other and establish a simple frequency distribution. On the other hand, two other variables that are not in the cross-classification matrix must be taken into account due to their relevancy and effect in a household trip generation. The number of people in the household and the total number of miles traveled in the household are factors that have a significant impact on the *hhldtrips* variable as later demonstrated in Figure 14 and 15.

Table 11. Cross classification matrix between hhldtrip and hhsize.

hhsize/ hhldtrips	1	2	3	4	Total
1	1	1	0	2	4
2	14	11	2	2	29
3	8	4	1	1	14
4	11	10	3	2	26
5	3	9	4	1	17
6	11	7	4	3	25
7	4	8	4	2	18
8	2	19	6	2	29
9	0	4	2	2	8
10	0	11	4	1	16
11	2	7	2	1	12
12	0	7	3	2	12
13	0	3	1	4	8
14	0	6	3	3	12
15	0	5	3	2	10
16	1	1	0	5	7
17	0	1	1	4	6
18	0	1	2	4	7
19	0	0	1	4	5
20	0	2	1	3	6
21	0	0	1	4	5
22	0	0	1	4	5
23	0	0	0	1	1
24	0	0	0	5	5
25	0	0	0	2	2
29	0	0	0	2	2
31	0	0	1	0	1
32	0	0	0	1	1
34	0	0	0	1	1
41	0	0	0	2	2
49	0	0	0	1	1
Total	57	117	50	73	297

Table 12. Cross classification matrix between hhldtrip and wrkcount.

wrkcount/ hhldtrips	0	1	2	3	Total
1	2	0	2	0	4
2	10	12	5	2	29
3	5	6	3	0	14
4	9	11	5	1	26
5	9	5	3	0	17
6	11	9	5	0	25
7	2	7	7	2	18
8	9	6	13	1	29
9	1	2	4	1	8
10	3	4	8	1	16
11	2	7	3	0	12
12	3	3	6	0	12
13	0	2	6	0	8
14	6	2	1	3	12
15	1	1	6	2	10
16	1	2	4	0	7
17	0	1	4	1	6
18	0	2	4	1	7
19	0	2	2	1	5
20	1	1	2	2	6
21	0	2	3	0	5
22	1	3	1	0	5
23	0	1	0	0	1
24	0	2	2	1	5
25	0	0	2	0	2
29	0	1	1	0	2
31	0	1	0	0	1
32	0	0	0	1	1
34	0	1	0	0	1
41	0	1	0	1	2
49	0	1	0	0	1
Total	76	98	102	21	297

Table x. Cross classification matrix between hhldtrip and hhvehcnt.

hhvehcnt/ hhldtrips	0	1	2	3	4	Total
1	0	2	1	1	0	4
2	3	11	14	0	1	29
3	2	7	4	1	0	14
4	4	17	3	2	0	26
5	1	8	7	1	0	17
6	2	12	7	1	3	25
7	0	6	7	3	2	18
8	1	8	17	0	3	29
9	0	1	3	4	0	8
10	0	3	9	4	0	16
11	0	4	6	2	0	12
12	1	4	5	2	0	12
13	0	0	7	0	1	8
14	0	4	5	1	2	12
15	0	2	4	1	3	10
16	0	1	3	3	0	7
17	0	1	4	0	1	6
18	0	1	3	2	1	7
19	0	2	1	2	0	5
20	0	0	3	3	0	6
21	0	0	3	2	0	5
22	0	0	5	0	0	5
23	0	0	0	1	0	1
24	0	1	2	1	1	5
25	0	0	2	0	0	2
29	0	1	0	1	0	2
31	0	1	0	0	0	1
32	0	0	0	0	1	1
34	0	0	1	0	0	1
41	0	0	1	1	0	2
49	0	0	0	1	0	1
Total	14	97	127	40	19	297

The best approach to understand the relationship of the explained variable and the multiple explanatory variables is to undertake multiple linear regression analysis. However, there are certain assumptions that need to be verified in order to guarantee the reasonableness of the models generated, such as normality, homoscedasticity, and linearity. All necessary aspects of the data were evaluated and the assumptions were respected. The dependent variable *hhldtrips* and *pertrips* given all the variables are normally distributed, the observations are independent, there is no perfect collinearity, E (E / x1, x2, x2, ..., xn) = 0, and there is homoskedaeticity.

Due to the format of categorical variables in both datasets, it was essential to develop a series of dummy variables that enabled their usage in the regression attempts. All the commands used in this report were implemented in the statistical application software package STATA and its do-files can be found in the Appendix. The process of exploring possible regression models is based on several attempts to find the best subsets of variables that affect the regressand variable. Two multiple linear regression models were estimated for *hhldtrips* and *pertrips*. The process of defining the variables used in the model took into consideration if the p-values were bellow 0.05, if the f-test was zero, and the coefficient of determination, which represents the variation explained by the model. The higher the R^2 is, the best the model explains the linear relation.

The model in Figure 1 used the variables *trpmiles*, *hhsize*, *highMidIncome*, *numadlt*, and *drvrcnt* to explain *hhldtrips*. For one unit change in each variable, the dependent variable will change by its coefficient. For example, for one unit change in the number of people in the household, the number of household trip will increase by 3.19 units. The same concept can be used to explain negative relation, as for one unit change in number of adults in a household, the *hhldtrip* decreases by 2.18 units. The ratio of variation explained by the model is 0.44. On the other hand, Figure 2 describes a model where only the variable *trpmiles*, *hhsize*, and *highMidIncome* are used and its coefficient of determination is 0.43. Thus, we have enough evidence to affirm that the second model explains the variation in a more practical way. In both scenarios, having a income between $50,000 and $69,999 would increase the household trips by 2.02 and 2.15 respectively.

First model: *hhldtrip* = 1.96 + 0.04(*trpmiles*) + 3.07(*hhsize*)
2.02(*highMidIncome*) - 2.18(*numadlt*) + 1.55(*drvcnt*)

Second model: *hhldtrip* = 0.965 + 0.004(*trpmiles*)
3.07(*hhsize*) + 2.15(*highMidIncome*)

Source	SS	df	MS			
				Number of obs	=	297
				F(5, 291)	=	46.99
Model	7181.03475	5	1436.20695	Prob > F	=	0.0000
Residual	8894.62855	291	30.5657339	R-squared	=	0.4467
				Adj R-squared	=	0.4372
Total	16075.6633	296	54.3096733	Root MSE	=	5.5286

hhldtrips	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
trpmiles	.0043225	.00122	3.54	0.000	.0019214	.0067235
hhsize	3.196176	.2902052	11.01	0.000	2.625009	3.767344
highMidIncome	2.025511	.8472747	2.39	0.017	.3579475	3.693074
numadlt	-2.188817	.8708431	-2.51	0.012	-3.902767	-.4748681
drvrcnt	1.555917	.7039404	2.21	0.028	.1704568	2.941377
_cons	1.961293	.9906506	1.98	0.049	.0115448	3.911042

Figure 1. First multiple regression model on *hhldtrips*.

Source	SS	df	MS			
				Number of obs	=	297
				F(3, 293)	=	74.86
Model	6975.19963	3	2325.06654	Prob > F	=	0.0000
Residual	9100.46367	293	31.059603	R-squared	=	0.4339
				Adj R-squared	=	0.4281
Total	16075.6633	296	54.3096733	Root MSE	=	5.5731

hhldtrips	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
trpmiles	.0044473	.0012285	3.62	0.000	.0020296	.006865
hhsize	3.070986	.2280238	13.47	0.000	2.622213	3.519758
highMidIncome	2.152489	.8430955	2.55	0.011	.4931987	3.81178
_cons	.9653369	.6906584	1.40	0.163	-.3939434	2.324617

Figure 2. Second multiple regression model on *hhldtrips*.

The following figures explore some characteristics of the person trip file. Figure 3 is the first model generated and explains the relationship between the dependent variable *pertrips* and the independent variables *male*, *lessThanHS*, *lowIncome*, *isDriver*. There is a poor fit in the linear relation since the coefficient of determination found was 0.0468. Besides *isDriver*, all the coefficients of the regression equation are negative, demonstrating the negative relationship with the number of person trips. In other words, if an observation describes a man, who has a low income and less than a high school diploma, the person trips would fall according to the corresponding coefficients.

On the other hand, the second model uses some antagonists to generate the analysis. We can observe the variables *female*, *greaterThanHS*, *highMidIncome*, and *isDriver* in the second model. The coefficient of determination is also low, 0.0484. In both scenarios, being a driver is a significant factor for a trip generation, since it would increase the *pertrip* in 0.80 and 0.90 respectively.

First model: *pertrips* = 4.34 -0.36(*male*) -0.83(*leassThanHS*) - 0.62(*lowIncome*) + 0.80(*isDriver*)

Second model: *pertrips* = 3.72 + *0.90*(*isDriver*) + 0.40(*female*) - 0.58(*greaterThanHS*) +0.59(*highMidIncome*)

```
. regress pertrips male lessThanHS lowIncome isDriver

      Source |       SS           df       MS      Number of obs   =       648
-------------+----------------------------------   F(4, 643)       =      7.89
       Model |  173.709317         4  43.4273293   Prob > F        =    0.0000
    Residual |   3537.9697       643  5.50228568   R-squared       =    0.0468
-------------+----------------------------------   Adj R-squared   =    0.0409
       Total |  3711.67901       647  5.73675272   Root MSE        =    2.3457

------------------------------------------------------------------------------
    pertrips |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -.3696064   .1851616    -2.00   0.046     -.733201   -.0060119
  lessThanHS |  -.8396789   .3421401    -2.45   0.014    -1.511526    -.167832
   lowIncome |  -.6271644   .2774494    -2.26   0.024    -1.171981   -.0823481
    isDriver |   .8037822   .2074073     3.88   0.000     .3965047     1.21106
       _cons |   4.349459   .2072625    20.99   0.000     3.942465    4.756452
------------------------------------------------------------------------------
```

Figure 3. First multiple regression model on pertrips.

```
. regress pertrips isDriver female greaterThanHS highMidIncome

      Source |       SS           df       MS      Number of obs   =       648
-------------+----------------------------------   F(4, 643)       =      8.17
       Model |  179.585834         4  44.8964586   Prob > F        =    0.0000
    Residual |  3532.09318       643  5.49314647   R-squared       =    0.0484
-------------+----------------------------------   Adj R-squared   =    0.0425
       Total |  3711.67901       647  5.73675272   Root MSE        =    2.3437

-------------------------------------------------------------------------------
     pertrips |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
     isDriver |   .9040963   .2103391     4.30   0.000     .4910618    1.317131
       female |   .4007624   .1848287     2.17   0.031     .0378217    .7637032
greaterThanHS |  -.5863476   .2377304    -2.47   0.014    -1.053169   -.1195257
highMidIncome |   .5983316   .2222811     2.69   0.007     .1618471    1.034816
        _cons |    3.72583   .2027348    18.38   0.000     3.327727    4.123932
-------------------------------------------------------------------------------
```

Figure 4. Second multiple regression model on pertrips.

Two additional models were performed to estimate a person trip linear regression models for adult males and females separately. With these conditions set, we were able to comprehend the differences in the elements that have a direct impact on the number of person trips between genders. As we can observe in Figure 5, *isNotWorker* and *drvrcnt* have a positive impact on *pertrips* for females over 18 years old. The *numadlt* has a negative impact. However, the Rˆ2 in this model is low, in accordance with the other models generated by the person trip dataset. For males over 18 years old, there were not many variables that have a P-value smaller than 0.05. Being a driver seems to have a positive impact, and having a degree higher than high school seems to have a negative impact in the *pertrips* variable, as we can see in Figure 6. The coefficient of determination is 0.0043.

First model: *pertrips* = 4.64 + 0.70(*isNotWorker*) -0.83(*numadlt*) + 0.74(*drvrcnt*)

Second model: *pertrips* = 3.81 +1.42(*isDriver*) -0.72(*greaterThanHS)*

```
. regress pertrips isNotWorker numadlt drvrcnt if r_sex == 1 & r_age >= 18

      Source |       SS           df       MS      Number of obs   =       225
-------------+----------------------------------   F(3, 221)       =      3.29
       Model |  50.8239429         3  16.9413143   Prob > F        =    0.0216
    Residual |  1139.55828       221   5.1563723   R-squared       =    0.0427
-------------+----------------------------------   Adj R-squared   =    0.0297
       Total |  1190.38222       224  5.31420635   Root MSE        =    2.2708

------------------------------------------------------------------------------
    pertrips |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 isNotWorker |   .7023268   .3470338     2.02   0.044     .0184077    1.386246
     numadlt |  -.8321624   .3658663    -2.27   0.024    -1.553196   -.1111292
     drvrcnt |   .7411575   .3001676     2.47   0.014     .1496003    1.332715
       _cons |   4.647273   .5315704     8.74   0.000     3.599677    5.694869
------------------------------------------------------------------------------
```

Figure 5. Multiple regression model on *pertrips* for females over 18 years old.

```
. regress pertrips isDriver greaterThanHS if r_sex == 2 & r_age >= 18

      Source |       SS           df       MS      Number of obs   =       251
-------------+----------------------------------   F(2, 248)       =      5.57
       Model |  76.5519458         2  38.2759729   Prob > F        =    0.0043
    Residual |  1703.95801       248  6.87079845   R-squared       =    0.0430
-------------+----------------------------------   Adj R-squared   =    0.0353
       Total |  1780.50996       250  7.12203984   Root MSE        =    2.6212

-------------------------------------------------------------------------------
     pertrips |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
     isDriver |   1.420676   .6048506     2.35   0.020     .2293767    2.611975
greaterThanHS |   -.729753   .3687654    -1.98   0.049    -1.456064   -.0034415
        _cons |   3.810823   .6037324     6.31   0.000     2.621726     4.99992
-------------------------------------------------------------------------------
```

Figure 6. Multiple regression model on *pertrips* for males over 18 years old.

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(19)     =     91.43
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |      91.43     19    0.0000
            Skewness |      13.40      5    0.0199
            Kurtosis |       3.34      1    0.0676
---------------------+-----------------------------
               Total |     108.17     25    0.0000
---------------------------------------------------
```

Figure 7. White's test for homoskedasticity for first model of hhldtrips.

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(8)      =     72.58
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |      72.58      8    0.0000
            Skewness |       5.90      3    0.1168
            Kurtosis |       3.57      1    0.0589
---------------------+-----------------------------
               Total |      82.04     12    0.0000
---------------------------------------------------
```

Figure 8. White's test for homoskedasticity for second model of hhldtrips.

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(10)     =     13.40
         Prob > chi2  =    0.2020

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |      13.40     10    0.2020
            Skewness |      19.74      4    0.0006
            Kurtosis |       3.76      1    0.0525
---------------------+-----------------------------
               Total |      36.91     15    0.0013
---------------------------------------------------
```

Figure 9. White's test for homoskedasticity for first model of pertrips.

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(10)     =     16.71
         Prob > chi2  =    0.0810

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |      16.71     10    0.0810
            Skewness |      20.54      4    0.0004
            Kurtosis |       4.36      1    0.0368
---------------------+-----------------------------
               Total |      41.61     15    0.0003
---------------------------------------------------
```

Figure 10. White's test for homoskedasticity for second model of pertrips.

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(8)      =      3.12
         Prob > chi2  =    0.9263

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |       3.12      8    0.9263
            Skewness |      13.35      3    0.0039
            Kurtosis |       1.18      1    0.2774
---------------------+-----------------------------
               Total |      17.65     12    0.1266
---------------------------------------------------
```

Figure 11. White's test for homoskedasticity for model on *pertrips* for females over 18 years old.

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(3)      =      3.03
         Prob > chi2  =    0.3864

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |       3.03      3    0.3864
            Skewness |       9.80      2    0.0074
            Kurtosis |       3.33      1    0.0679
---------------------+-----------------------------
               Total |      16.17      6    0.0129
---------------------------------------------------
```

Figure 12. White's test for homoskedasticity for model on *pertrips* for males over 18 years old.

The White's Test is used to understand the heteroscedastic errors in multiple regression models analysis. Checking wether the variance of the error is constant is a critical characteristic that helps to define if the conditions are satisfied. The null hypothesis represents the homoskedasticity of the data, and the alternative hypothesis the unrestricted heteroskedasticty. As we can observe in Figure 7 and Figure 8, the p-value is less than 0.05, therefore we have enough evidence to reject the homeskedasticity hypothesis. On the other hand, Figure 9 through 12 shows a p-value tiger than 0.05, thus, we fail to reject the null hypothesis.

Conclusion

Throughout this report, six multiple regression models were generated to comprehend the relationship between the regressant variables *pertrips* and *hhldtrips* and the best regressor variables of the NHTS Phoenix-Mesa sub-sample. Even though the coefficient of determination is considered low in all scenarios, we can identify some relevant variables in the datasets. The household size is one of the most impactful element and the household income between $50,000 and $69,999 is the most significant range affecting household trips.

For the person trips, no significant model was discovered due to the extremely low R^2. However, the negative coefficients on the first regression model is meaningful, since some of the elements are the opposite dummy variable of the second model. Being a driver, being female, and having an income between $50,000 and $69,999 are characteristics that seems to have an impact on persons trip. Finally, when analyzing the same dataset but for females over 18 years old, *isNotWorker*, *numadlt*, and *drvrcnt* were the chosen elements to include in the model and for males over 18 years old, *isDriver* and *greaterThanHS* were chosen. There is a poor fit of the data in the model in both cases.

Appendix

Do-file 1:

```
import excel "/Users/miguelamaral/Downloads/persontrips.xlsx", sheet("persontrips") firstrow
tabulate driver devrcnt
tabulate driver drvrcnt
tabulate driver
tabulate drvrcnt
tabulate educ
tabulate hhincttl
summarize hhsize drvrcnt r_age hhsize pertrips
tabulate worker
summarize hhsize drvrcnt r_age numadlt pertrips
tabulate r_sex
tabulate homeown
gen isDriver = cond(driver == 1, 1, 0)
gen isNotDriver = cond(driver == 2, 1, 0)
gen isWorker = cond(worker == 1, 1, 0)
gen isNotWorker = cond(worker == 2, 1, 0)
gen lessThanHS = cond(educ == 1, 1, 0)
gen greaterThanHS = cond(educ == 2, 1, 0)
gen lowIncome = cond(hhincttl == 1 | hhincttl == 2 | hhincttl == 3 | hhincttl == 4 | hhincttl == 5, 1, 0)
gen lowMidIncome = cond(hhincttl == 6 | hhincttl == 7 | hhincttl == 8 | hhincttl == 9 | hhincttl == 10, 1, 0)
gen highMidIncome = cond(hhincttl == 11 | hhincttl == 12 | hhincttl == 13 | hhincttl == 14, 1, 0)
gen highIncome = cond(hhincttl == 15 | hhincttl == 16 | hhincttl == 17 | hhincttl == 18, 1, 0)
gen male = cond(r_sex == 1, 1, 0)
gen female = cond(r_sex == 2, 1, 0)
gen ownHome = cond(homeown == 1, 1, 0)
gen rentHome = cond(homeown == 2, 1, 0)
regress pertrips isDriver lessThanHS highMidIncome
regress pertrips r_age highIncome highMidIncome lowIncome lowMidIncome
regress pertrips male female r_age drvrcnt
regress r_age highMidIncome lessThanHS
regress pertrips r_age highMidIncome lessThanHS
regress pertrips r_age highMidIncome greaterThanHS
regress pertrips male ownHome r_age numadlt highIncome lessThanHS isWorker isDriver
regress pertrips female r_age numadlt highMidIncome lessThanHS isWorker isDriver
regress pertrips female numadlt highMidIncome lessThanHS isDriver lowIncome
 regress pertrips female highMidIncome lessThanHS isDriver lowIncome
 regress pertrips isDriver isNotDriver isNotWorker lowIncome highMidIncome
regress pertrips isDriver lowIncome greaterThanHS
regress pertrips lessThanHS isDriver lowIncome female rentHome
regress pertrips isDriver female greaterThanHS highMidIncome
estat imtest, white
regress r_age male lessThanHS lowIncome isDriver
regress pertrips r_age male lessThanHS lowIncome isDriver
regress pertrips male lessThanHS lowIncome isDriver
estat imtest, white
regress pertrips female lessThanHS lowIncome isDriver
regress pertrips if r_sex == 1 & r_age >= 18
regress pertrips if r_sex == 2 & r_age >= 18
 regress pertrips isDriver greaterThanHS lessThanHS highIncome highMidIncome if r_sex == 1 & r_age
>= 18
```

```
regress pertrips isDriver lessThanHS lowIncome lowMidIncome ownHome rentHome  if r_sex == 1 &
r_age >= 18
regress pertrips isDriver lessThanHS isNotWorker isWorker isNotDriver if r_sex == 1 & r_age >= 18
regress pertrips lessThanHS isNotWorker isNotDriver if r_sex == 1 & r_age >= 18
regress pertrips lessThanHS isNotWorker numadlt drvrcnt hhsize if r_sex == 1 & r_age >= 18
regress pertrips lessThanHS isNotWorker numadlt if r_sex == 1 & r_age >= 18
regress pertrips lessThanHS isNotWorker drvrcnt if r_sex == 1 & r_age >= 18
regress pertrips lessThanHS numadlt drvrcnt if r_sex == 1 & r_age >= 18
regress pertrips isNotWorker numadlt drvrcnt if r_sex == 1 & r_age >= 18
estat imtest, white
regress pertrips isNotDriver rentHome lowIncome lowMidIncome lessThanHS if r_sex == 1 & r_age >=
18
regress pertrips isNotDriver rentHome lowIncome lowMidIncome lessThanHS if r_sex == 2 & r_age >=
18
regress pertrips isDriver ownHome  highMidIncome highIncome greaterThanHS  if r_sex == 2 & r_age >=
18
regress pertrips isDriver greaterThanHS isNotWorker numadlt drvrcnt  if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS isWorker drvrcnt hhsize if r_sex == 2 & r_age >= 18
regress pertrips isNotDriver isWorker greaterThanHS isNotWorker drvrcnt  if r_sex == 2 & r_age >= 18
regress pertrips isDriver isWorker greaterThanHS if r_sex == 2 & r_age >= 18
regress pertrips isDriver isWorker greaterThanHS lowIncome lowMidIncome  if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS highIncome highMidIncome if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS isWorker if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS isNotWorker  if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS rentHome  if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS ownHome if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS hhsize if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS drvrcnt driver if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS drvrcnt if r_sex == 2 & r_age >= 18
regress pertrips isDriver greaterThanHS if r_sex == 2 & r_age >= 18
estat imtest, white
```

Do-file 2:

```
import excel "/Users/miguelamaral/Downloads/Hhldtrips.xlsx", sheet("Hhldtrips") firstrow
tabulate homeown
summarize hhvehcnt hhsize drvrcnt wrkcount numadlt trpmiles
tabulate hhincttl
gen ownHome = cond(homeown == 1, 1, 0)
gen rentHome = cond(homeown == 2, 1, 0)
gen lowIncome = cond(hhincttl == 1 | hhincttl == 2 | hhincttl == 3 | hhincttl == 4 | hhincttl == 5, 1, 0)
gen lowMidIncome = cond(hhincttl == 6 | hhincttl == 7 | hhincttl == 8 | hhincttl == 9 | hhincttl == 10, 1, 0)
gen highMidIncome = cond(hhincttl == 11 | hhincttl == 12 | hhincttl == 13 | hhincttl == 14, 1, 0)
gen highIncome = cond(hhincttl == 15 | hhincttl == 16 | hhincttl == 17 | hhincttl == 18, 1, 0)
regress hhldtrips wrkcount trpmiles numadlt
regress hhldtrips wrkcount trpmiles numadlt homeown hhvehcnt
regress hhldtrips trpmiles numadlt hhsize drvrcnt rentHome
regress hhldtrips trpmiles hhsize drvrcnt ownHome lowIncome lowMidIncome
regress hhldtrips trpmiles hhsize ownHome highIncome highMidIncome rentHome numadlt
regress hhldtrips trpmiles hhsize ownHome highMidIncome rentHome drvrcnt
regress hhldtrips trpmiles hhsize highMidIncome numadlt drvrcnt
estat imtest, white
```

```
regress hhldtrips trpmiles hhsize highMidIncome
regress hhldtrips trpmiles hhsize numadlt
estat imtest, white
regress hhldtrips trpmiles hhsize highMidIncome numadlt drvrcnt rentHome lowIncome ownHome
replace hhsize = 4 if hhsize == 5 | hhsize == 6 | hhsize == 7 | hhsize == 8 | hhsize == 9 | hhsize == 10
tabulate hhldtrips hhsize
replace hhvehcnt = 4 if hhvehcnt == 5 | hhvehcnt == 6 | hhvehcnt == 7
tabulate hhldtrip hhvehcnt
replace wrkcount = 3 if wrkcount == 5 | wrkcount == 4
tabulate hhldtrip wrkcount
```