WMATA's Land Use-Ridership Model

March 2015

Office of Planning

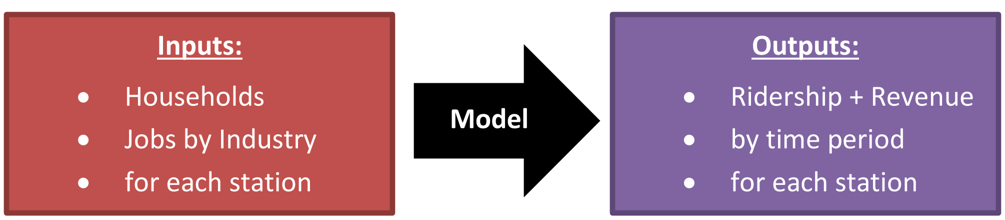
What Is a Land Use-Ridership Model?

It's a tool that the Planning Office has built that will predict changes in Metrorail ridership as a result of land use changes in the station area. If we build a new apartment building next to a Metrorail station, for instance, how much ridership will WMATA realize?

This tool is based on a solid understanding of the link between land use and the rail ridership we see today. To build this, we analyzed what you can actually walk to from each station, assembled detailed information about land uses and densities in those areas (households, jobs by industry type), and also controlled for other, non-land-use factors that shape ridership – like network accessibility. More details below.

What Can I Use It For?

Most immediately, the Land Use-Ridership model can predict changes in ridership and revenue as a result of changes in land use.



The model can also answer questions such as:

- If we build an office building at Station X or Station Y, which generates more ridership?
- For a given amount of commercial space near a station, does office or retail generate more ridership, and at what times of day?
- What kinds of development produce ridership at off-peak times?
- How much density would be required to generate $X of fare revenue?

The Good: Features and Strengths

The three key strengths of this model are:

- It accounts for a variety of factors that explain ridership differently at different stations, notably network accessibility, neighborhood socioeconomics, and the quality of rail service.

- It's based on a **comprehensive** look at how all stations perform right now. The model looks at data on *all* land uses and ridership at *all* Metrorail stations. No sampling, or using national averages.
- The model is **station-specific**, meaning that it adjusts the ridership forecasts for the station, utilizing what we know about the station now. For example, we generally know that the number of jobs at a station determines PM Peak entries. But we also know that rate is higher for stations with higher access to households, so the model yields a higher forecast for a station with good household access.

This last point is critical, and is discussed further below.

The Bad: Drawbacks and Caveats

This model is not a 100% answer, but it's one of the best estimates available. The supporting modeling achieves R^2 values in the range of 0.7 to 0.9, meaning that the modeling explains only **70-90%** of the difference in ridership across stations, stronger for peak periods and weaker for off-peak. Because it's based on multivariate regression models, it can't include the effects of every factor if they are collinear.

This is the first phase of the model, so for now it:

- Covers only Metrorail so far (not bus),
- Covers only walk and bike ridership, since riders who arrive by bus or car are coming from farther away and have little connection to the land uses right around the station, and
- Predicts where riders will enter, but not exit. In other words, it will predict ridership, but won't predict where the new riders will go. So for revenue estimates, it is sometimes necessary to double the revenue assuming round-trips.

Where Did This Model Come From?

The Land Use-Ridership model is based on four multivariate regressions, predicting walk ridership by time of day (AM Peak, Midday, PM Peak, Evening) as a function of the land uses in the station's walkable area, characteristics of that station's role in the Metrorail network, and other factors.

Figure 1. General land use-ridership model structure

The walkable area is defined as a half-mile walk along a road network, so we account for barriers like highways and bridges. The size of the resulting "[walk sheds]" differ significantly across stations, from a large shed at say Ballston, and a small shed at Cheverly, where Route 50 is a barrier. The half-mile cutoff is a bit longer than the median actual walk distance reported by our riders in the 2012 Metrorail Passenger Survey.

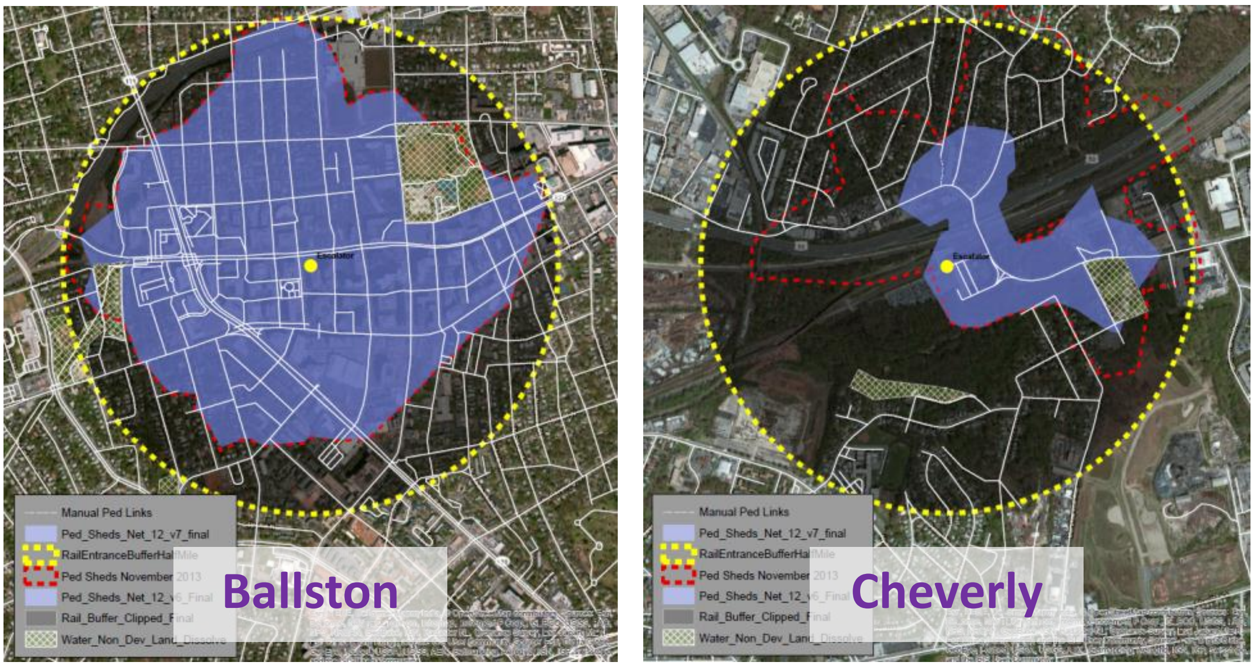


Figure 2. Half-mile walkable area from two sample stations

For each station and its walk shed, we tested the following kinds of factors:

- Number of households, and number of jobs, by industry type (NAICS code)
- Demographics like median income
- Built environment variables like block density, WalkScore, land use diversity
- Accessibility to jobs and households scores via Metrorail from that station
- Metrorail service characteristics like trains per hour, transit connectivity index
- Relative competitiveness of Metrorail vs. driving (access via rail vs. auto drive times in congested conditions, cost of private parking)
- Interactive terms between households, jobs, and other factors

We enlisted the help of University of Maryland's National Center for Smart Growth at this point, to help with the technical aspects of the statistics, for datasets, and their prior experience with this kind of analysis. We used all these variables in multivariate regressions to predict walk ridership, using September 2013 ridership data from the fare system, and the 2012 Metrorail Passenger Survey for access mode. The resulting coefficients are applied in the model to predict ridership.

Revenue impacts are estimated using the October 2014 average fare (peak and off-peak) from the station. Essentially, this means we assume that new riders will take trips of similar length and fare as current riders.

What Does Network Accessibility Mean?

One major innovation of this model is from our demonstrating that *accessibility* helps explain a great deal of why people choose to ride Metrorail, so we should consider it when we predict ridership. But what does accessibility mean?

Accessibility means how much useful stuff – households, jobs, stores, etc. – you can *get to* via Metrorail within a certain amount of time from a given station – in this model, 30 minutes. We used an arbitrary cutoff to start, but we can improve this using a decay function or other methods in the future. This measures the *value* of the rail network to a rider, and it turns out to strongly help predict ridership.

Consider a commuter who lives adjacent to Crystal City, compared to Greenbelt. In 30 minutes from Crystal City, a rider could reach 42 other Metrorail stations; from Greenbelt 13. And, most importantly, the *jobs at* those 42 stations from Crystal City total over 1.1 million (including downtown DC), over ten times more than from Greenbelt. The resident at Crystal City is much more likely take Metrorail because its jobs access is higher - there's simply a higher likelihood that their job will be metro-accessible.

The same phenomenon holds in reverse, too: employers located near stations with better access to households better attract riders via Metrorail.

From Crystal City: **From Greenbelt:**

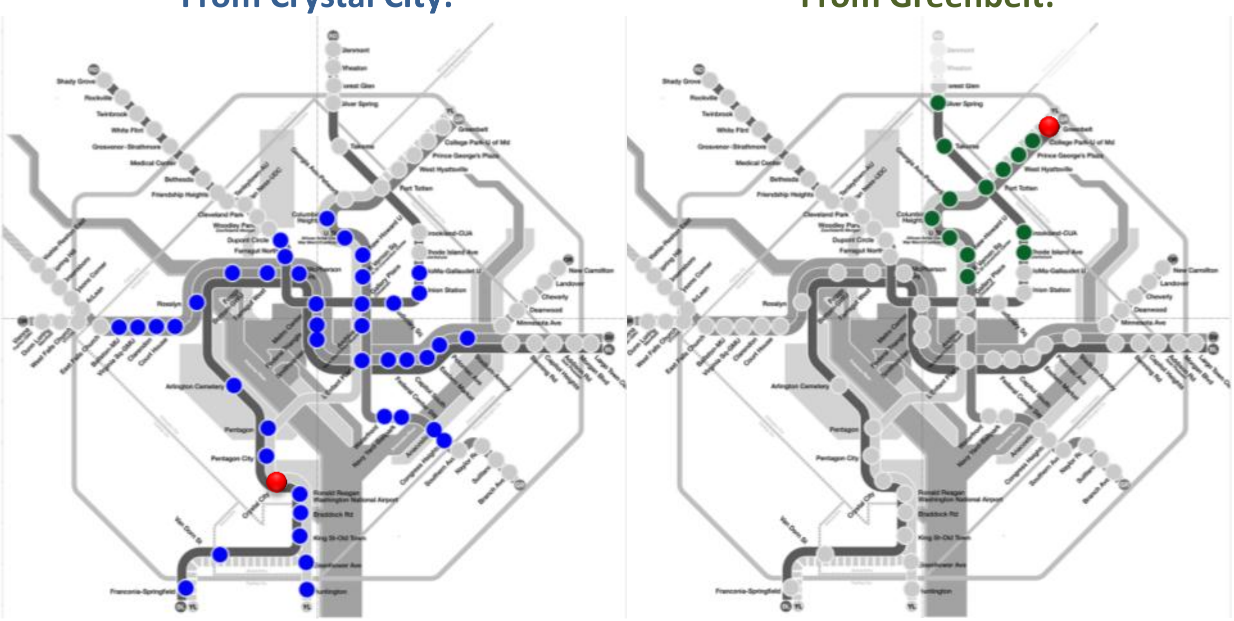


Figure 3. Network Accessibility: Stations reachable within 30 minutes from two sample stations

Accessibility scores significantly help explain, for the same land use change, variations in ridership per household (or job) across stations today. This Land Use-Ridership model applies each station's accessibility score to produce a ridership estimate tailored to the individual station.

Why Is Land Use Important?

Because the land use around Metrorail stations is a huge deciding factor in why people take Metrorail in the first place, and this translates into big impacts on our costs and revenues. Land use helps explain why walk ridership is over fifteen times higher at Columbia Heights than at Cheverly, for instance. Bethesda's mix of jobs and households helps explain why that station is utilized evenly in the morning and the evening, and why we get much more bang for the buck out of that station compared to commuter-only stations.

Conversely, big increases in density can also increase ridership enough that it strains a station's vertical circulation capacity (elevators, escalators, faregates), and can trigger the need to add more capacity.

In short, land use is a huge component of Metro's cost and revenue structure.

How Is This Different From What We've Done Before?

In the past, we have begun with the number of jobs and people in a land use change, and applied a trip-generation rate for transit. The numbers are largely based on WMATA's 2005 Development-Related Ridership Survey, where Metro conducted in-field surveys of travel patterns at a sample of sites around the Metrorail system.

This methodology has advantages:

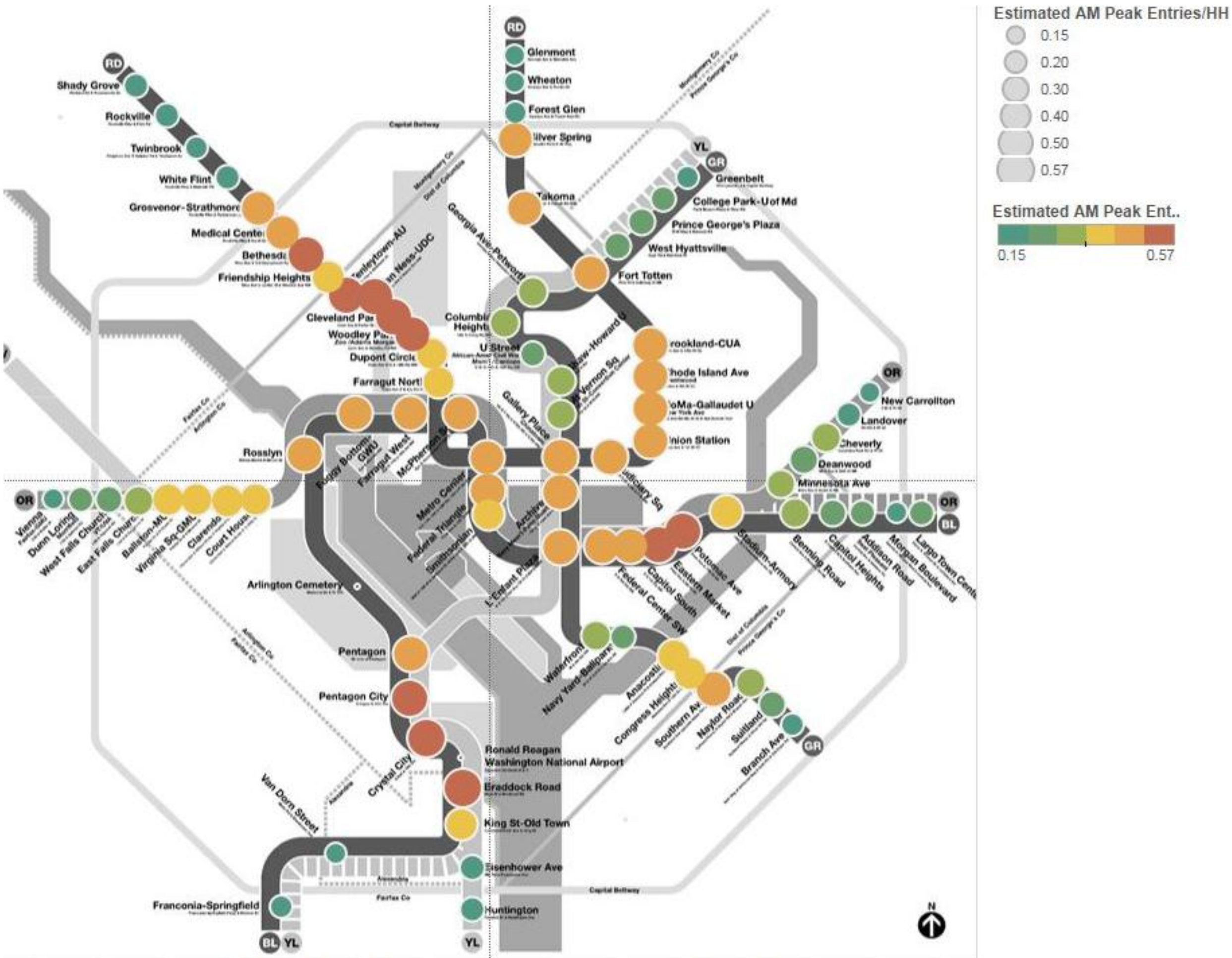
- It's based on original survey work which more closely measures the link between travelers and ridership, rather than the "desk exercise" of comparing raw land use data to raw faregate counts.
- It considered the distance a building was from Metrorail, in quarter-mile increments
- It asks about all modes like auto, including other transit modes, like Metrobus, commuter rail, and walking. In this way it can give a richer picture of overall travel characteristics from a development, rather than simply the number of Metrorail trips generated.

In other ways, the Land Use-Ridership model has advantages:

- It uses more current (2013) data than the 2005 study
- It uses a 100% sample of all stations, since all data sources were readily available without surveying
- It leverages more precise data about each station area that can explain ridership generation above and beyond the number of households or jobs – factors like accessibility, neighborhood demographics, and rail service.

The findings between the two tools are broadly consistent, however, and we'd recommend using both tools when estimating ridership from a real estate development. Both are different, equally valid ways of answering similar questions.

Sample Results – AM Peak



TECHNICAL APPENDIX A – REGRESSION SPECIFICATIONS

The Land Use-Ridership model uses the following final regression specifications to determine the coefficients applied. Regressions were conducted in Stata IC 13.1, and tested for heteroskedasticity using the Breusch-Pagan / Cook-Weisberg test for heteroskedasticity.

Technical notes:

- A single model specification was used for AM Peak Entries, Midday, and Evening. For PM Peak, stations were divided into three tiers based on job density, and a separate regression estimated for each tier.
- The specification for PM Peak Entries actually uses AM Peak Exits as a proxy dependent variable, because the AM Peak is a "cleaner" commute market to model.
- Walk sheds used were non-overlapping, meaning that if a household was within more than one walk shed, it was assigned to the nearest station. This avoids double-counting trip generators.
- Number of jobs is a proxy for the general level of activity, particularly in the off-peak regressions.
- Regressions were estimated using a variety of independent variables; in the end, final models applied were "parsimonious" where all statistically-insignificant are dropped out.
- Silver Line stations are not included in the data to generate the specifications (data on walk ridership was not available yet in 2014), but the model will estimate ridership changes at those stations.

Key Variables Used in Final Models

Variable Name	Description	Source
`Walk_Bike_Entries_AMPe~201`	Dependent variable. AM Peak Entries by walking and bicycling access. Ridership counts from average weekday in September 2013, multiplied by walk+bike access mode share from 2012 Metrorail Passenger Survey. Corresponding data for other time periods also used.	WMATA
`Households000050miles`	Number of households in the half-mile walk shed of a station, 2012. Block groups apportioned to walk sheds using area.	ESRI 2012 Demographics by Block Group
`HHsXJobsAccessRailvHighway`	Jobs Access (sum of all jobs [`jobshalf`] in walk sheds of stations that are reachable by Metrorail within 30 minutes) from the station, rail divided by highway. Interactive term between households and jobs access.	Transit: WMATA Highway: MWCOG
`MedianHHIncome`	Median income of block groups in station area	ESRI 2012
`HHsXGoodService`	Good service defined as combined 40 trains per hour in all directions in the AM rush hour; pre-Silver Line Metrorail schedule (September 2013). Interactive term between households and service quality.	WMATA
`IntersectionH`	Number of 3-way intersections in the station area walk shed; proxy for urban design	WMATA
`jobs_schools`	Number of jobs in the station area in the education industry (NAICS 21)	WMATA, ESRI
`jobs_nightsandweekends`	Number of jobs in the station area in the retail, restaurant, and entertainment industries (NAICS 44, 45, 71, 72)	WMATA, ESRI
`jobs_ninetofive`	Number of jobs in the station area in the office sector likely to have a 9am - 5pm schedule (NAICS 33, 51-56, 813, 92	
`HHsINWALKSHEDSOFSTATIONSWITH`	Household access. Sum of households (`Households000050miles`) in walk sheds of stations that are reachable by Metrorail within 30 minutes.	WMATA, ESRI
`jobshalf`	Number of jobs in the walk shed of a station	WMATA, ESRI
`jobsXHHAccess`	Interactive term between jobs (`jobshalf`) and household access (`HHsINWALKSHEDSOFSTATIONSWITH`)	WMATA, ESRI
`PrivateJobsLODES`	Number of private (non-governmental) jobs in the station's walk shed. Data available by block from the Census Bureau's 2011 LEHD LODES product.	2011 LEHD LODES, U.S. Census Bureau
`TPHPeakV2`	Trains per hour at the station in all directions in the AM rush hour; September 2013 Metrorail schedules.	

AM Peak Entries – Regression Results

```
      Source |       SS       df       MS              Number of obs =      84
-------------+------------------------------           F(  6,    77) =   67.60
       Model |  53575969.9     6  8929328.31           Prob > F      =  0.0000
    Residual |  10171122.3    77  132092.498           R-squared     =  0.8404
-------------+------------------------------           Adj R-squared =  0.8280
       Total |  63747092.2    83  768037.255           Root MSE      =  363.45

------------------------------------------------------------------------------
Walk_Bike_Entries_AMPe~201 |   Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------------+-------------------------------------------------
     Households000050miles | .1380333   .0520178     2.65   0.010     .0344526     .241614
HHsXJobsAccessRailvHighway | .2465355   .0597395     4.13   0.000      .127579     .365492
             MedianHHIncome |  .0051676   .0014789     3.49   0.001     .0022228    .0081124
            HHsXGoodService |   .157753   .0317532     4.97   0.000     .0945244    .2209816
              IntersectionH | -3.451063   2.784595    -1.24   0.219    -8.995899    2.093774
                      area1 | -296.3566   174.9304    -1.69   0.094    -644.6876    51.97444
                      _cons |  101.5836   115.2807     0.88   0.381    -127.9696    331.1369
------------------------------------------------------------------------------------------
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of Walk_Bike_Entries_AMPeak_Sept201

         chi2(1)      =     2.67
         Prob > chi2  =   0.1023
```

Midday Entries – Regression Results

```
. reg Walk_Bike_Entries_Midday_Sept201 jobs_schools jobs_nightsandweekends jobs_ninetofive
Households000050miles HHsINWALKSHEDSOFSTATIONSWITH MedianHHIncome

      Source |       SS       df       MS              Number of obs =      94
-------------+------------------------------           F(  6,    87) =   40.13
       Model |  50106568.5     6  8351094.75           Prob > F      =  0.0000
    Residual |  18102701.3    87  208077.026           R-squared     =  0.7346
-------------+------------------------------           Adj R-squared =  0.7163
       Total |  68209269.7    93  733433.008           Root MSE      =  456.15

----------------------------------------------------------------------------------------------
Walk_Bike_Entries_Midday~201 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------------+----------------------------------------------------------------
                jobs_schools |   .1480619   .0504321     2.94   0.004     .0478226    .2483012
      jobs_nightsandweekends |   .2117593   .0534826     3.96   0.000     .1054569    .3180617
             jobs_ninetofive |   .0298268   .0056849     5.25   0.000     .0185275    .0411262
       Households000050miles |   .0316094   .0287098     1.10   0.274    -.0254545    .0886733
HHsINWALKSHEDSOFSTATIONSWITH |   .0079576   .0013889     5.73   0.000      .005197    .0107182
              MedianHHIncome |  -.0010749   .0017486    -0.61   0.540    -.0045505    .0024007
                       _cons |  -59.62712   166.7146    -0.36   0.721    -390.9904    271.7361
----------------------------------------------------------------------------------------------
```

PM Peak Entries – Regression Results

GROUP 1 - Downtown CBD

```
reg Walk_Bike_Exits_AMPeak_Sept2013 jobshalf  if hiro_CBD0 == 1

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   45.00
       Model |   275999524     1   275999524           Prob > F      =  0.0000
    Residual |   116522500    19  6132763.18           R-squared     =  0.7031
-------------+------------------------------           Adj R-squared =  0.6875
       Total |   392522024    20  19626101.2           Root MSE      =  2476.4

------------------------------------------------------------------------------
Walk_Bi~2013 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    jobshalf |   .2781145    .041457     6.71   0.000     .1913441    .3648849
       _cons |   1493.753   909.2674     1.64   0.117    -409.3657    3396.871
------------------------------------------------------------------------------

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of Walk_Bike_Exits_AMPeak_Sept2013

         chi2(1)      =     2.13
         Prob > chi2  =   0.1440
```

GROUP 2 - LOW JOBS

```
. reg Walk_Bike_Exits_AMPeak_Sept2013 jobshalf TPHPeakV2 logMedianHHInc jobsXHHAccess if groupdumm
==2 & SHED_NAME ~="Suitland"

      Source |       SS       df       MS              Number of obs =      37
-------------+------------------------------           F(  4,    32) =   16.48
       Model |  838133.342     4  209533.336           Prob > F      =  0.0000
    Residual |  406928.252    32  12716.5079           R-squared     =  0.6732
-------------+------------------------------           Adj R-squared =  0.6323
       Total |  1245061.59    36  34585.0443           Root MSE      =  112.77

------------------------------------------------------------------------------------
Walk_Bike_Ex~2013 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------------
         jobshalf |   .1055366   .0557523     1.89   0.067    -.0080272    .2191004
        TPHPeakV2 |   3.648267   2.114945     1.72   0.094    -.6597354     7.95627
  logMedianHHIncome |  -82.10261   49.44629    -1.66   0.107    -182.8214    18.61619
    jobsXHHAccess |   1.14e-06   7.96e-07     1.43   0.163    -4.86e-07    2.76e-06
            _cons |   878.1165   532.9633     1.65   0.109    -207.4942    1963.727
```

GROUP 3 (MIXED JOBS)

```
. reg Walk_Bike_Exits_AMPeak_Sept2013 jobshalf  TPHPeakV2  jobsXHHAccess   if hiro_CBD0 == 0 &
jobs2500 ==0 & SHED_NAME ~= "Rosslyn"

      Source |       SS       df       MS              Number of obs =      24
-------------+------------------------------           F(  3,    20) =   47.58
       Model |  32304479.4     3  10768159.8           Prob > F      =  0.0000
    Residual |   4526505.2    20  226325.26            R-squared     =  0.8771
-------------+------------------------------           Adj R-squared =  0.8587
       Total |  36830984.6    23  1601347.16           Root MSE      =  475.74

------------------------------------------------------------------------------
Walk_Bik~2013 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     jobshalf |    .037676   .0166925     2.26   0.035     .0028561     .072496
    TPHPeakV2 |   53.72432   14.01354     3.83   0.001     24.49259    82.95604
jobsXHHAccess |   5.97e-07   1.79e-07     3.33   0.003     2.24e-07    9.71e-07
        _cons |  -1020.201   443.9406    -2.30   0.032    -1946.245   -94.15722
------------------------------------------------------------------------------
```

Evening Entries – Regression Results

```
. reg  Walk_Bike_Entries_Evening_Sept20  jobs_nightsandweekends  jobs_schools   PrivateJobsLODES
HHsINWALKSHEDSOFSTATIONSWITH

      Source |       SS       df       MS              Number of obs =      94
-------------+------------------------------           F(  4,    89) =   43.35
       Model |  69784107.3     4  17446026.8           Prob > F      =  0.0000
    Residual |    35816121    89  402428.326           R-squared     =  0.6608
-------------+------------------------------           Adj R-squared =  0.6456
       Total |   105600228    93  1135486.33           Root MSE      =  634.37

------------------------------------------------------------------------------------------------
Walk_Bike_Entries_Evening~20 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------------+------------------------------------------------------------------
      jobs_nightsandweekends |   .3600031   .0807379     4.46   0.000     .1995787    .5204275
                jobs_schools |   .2211178   .0716993     3.08   0.003     .0786528    .3635829
            PrivateJobsLODES |   .0289792   .0104458     2.77   0.007     .0082235    .0497348
HHsINWALKSHEDSOFSTATIONSWITH |   .0084478   .0017531     4.82   0.000     .0049644    .0119312
                       _cons |  -393.4991   128.8241    -3.05   0.003    -649.4699   -137.5283
------------------------------------------------------------------------------------------------
```