HADS custom variable definitions

* OWNRENT (comes from Tenure)
  + 0 – Rental
  + 1 – Owner
* COST06, COST08, COST12, COSTMED
  + Calculated cost to owner, this may be what we want to predict
* UTILITY
  + Is imputed for vacant units using monthly rent, structure type, region and tenure
* BURDEN
  + -1 for houses with no income, otherwise housing cost divided by monthly income
* Assisted housing
  + Overreported, use with caution

Variables that might be useful for prediction

ZINC2 – Household Income

IPOV – Poverty level income (according to the area)

ABLMED, ABL30, ABL50, ABL80 – Median income adjusted for bedrooms

ASSISTED – 0, 1, -9 – Are they receiving some kind of government assistance

FMR – Fair market rent – calculated by HUD as a factor of many things for a given county

LMED, L80, L50, L30 – Income levels of area

APLMED – Income adjusted for persons in household (AHS underestimates sometimes)

age1 – Age of head of household

BEDRMS

BUILT

OWNRENT

METRO3

REGION – Census Region

STATUS

STRUCTURETYPE / TYPE

ZADEQ – Adequacy of unit

ZSMHC – Monthly housing costs – we may have to delete this because it includes mortgage

NUNITS - # units of building

OTHERCOST – ZSMHC may be inconsistent, OTHERCOST is sum of insurance, land rent (not rent), and fees.

PER (# Of persons in household)

ROOMS (# Of Rooms in Unit)

TENURE (Owner/renter status of unit)

TOTSAL – Sum of salary income over all members of household

UTILITY (Monthly utility cost)

VALUE – Current market value (what we are trying to predict)

Coded Variables (taken almost verbatim from IV. A on the pdf):

AGE1

* -9, 0 – 93
* Group into buckets (13 - 33, 34 - 53, 54 - 73, 74 – 93)
  + Age1 (13-33)
  + Age2 (34 – 53)
  + Age3 (54 – 73)
  + Age4 (74 – 93)

Adequacy (ZADEQ):

* adequate, moderately inadequate, severely inadequate, vacant–no information
* We will use vacant-no information as our ‘0’variable

Bedrooms (BDRMS) #NOTE – while the pdf says this is coded, it appears it is not, there are 0 to 8 bedrooms in the document. It is possible bedrooms were coded for earlier years:

* PDF
  + Studio, 1, 2, 3, 4 or more
  + Studio will be our 0 variable
* Dataset
  + 0 – 8
  + 0 will be our 0 variable
  + OR we could make this continuous

Location #NOTE: doesn’t appear to be in dataset. There is a region variable which is for the census regions, but with 4 categories

* PDF
  + Central city, suburb, nonmetropolitan
  + Nonmetropolitan: 0
* Dataset
  + 1,2,3,4
  + 1 will be 0

METRO3: no ‘categorical’ notes in the pdf

* 1,2,3,4,5,9
* 9 will be 0

Structure Type:

* Single unit, 2-4 units, 5-19 units, 20-49 units, 50+ units, mobile homes, no info
* 1,2,3,4,5,6,-9
* No info (-9) will be 0
* Recoded from Type; we should not use both TYPE and STRUCTURETYPE

Tenure:

* Owner, renter
* Dataset has 1,2,3,-6
* -6 will be 0

Year Built

* After 201035, 2000-2009, 1990-1999, 1980-1989, 1960-1979, 1940-1959, Pre-  1940.
* Pre 1940 will be 0

Discard Variables:

BURDEN – is correlated to ‘Value’

CONTROL – The AHS control number. I believe this is just a private key

COSTX – anything cost related, because we can directly compute the fair market value with that

Anything FMR – these are all just recalculated percentages of FMR

ADEQ – Not in model, we use ZADEQ

FMR -

Anything that relates the housing price to another variable: affordability related to AMI and poverty income

Df Sum of Sq RSS AIC

<none> 5.2604e+15 2899832

+ B2000s 1 5.9216e+10 5.2603e+15 2899833

+ B90s 1 5.7624e+10 5.2603e+15 2899833

+ ABL80 1 2.6356e+09 5.2604e+15 2899834

- home\_age 1 2.5016e+11 5.2606e+15 2899835

- GL50 1 3.6025e+11 5.2607e+15 2899838

- GL30 1 4.4534e+11 5.2608e+15 2899840

- UTILITY 1 4.5800e+11 5.2608e+15 2899840

- ABL50 1 7.5904e+11 5.2611e+15 2899847

- ABL30 1 9.5004e+11 5.2613e+15 2899851

- IPOV 1 1.3075e+12 5.2617e+15 2899859

- B40s50s 1 1.5506e+12 5.2619e+15 2899865

- struc4 1 1.5670e+12 5.2619e+15 2899865

- struc2 1 1.7925e+12 5.2622e+15 2899870

- B80s 1 2.0516e+12 5.2624e+15 2899876

- struc3 1 2.9077e+12 5.2633e+15 2899895

- B2010s 1 4.3258e+12 5.2647e+15 2899927

- PER 1 4.6617e+12 5.2650e+15 2899935

- B60s70s 1 6.1716e+12 5.2665e+15 2899969

- struc5 1 6.2759e+12 5.2667e+15 2899971

- ABLMED 1 8.1848e+12 5.2686e+15 2900014

- GL80 1 8.9090e+12 5.2693e+15 2900030

- APLMED 1 9.1144e+12 5.2695e+15 2900035

- GLMED 1 1.0625e+13 5.2710e+15 2900069

- struc1 1 2.5589e+13 5.2860e+15 2900404

- TOTSAL 1 2.8874e+13 5.2893e+15 2900477

- ROOMS 1 4.1194e+13 5.3016e+15 2900752

- BEDRMS 1 5.1890e+13 5.3123e+15 2900991

- OTHERCOST 1 5.7748e+13 5.3181e+15 2901121

- AGE1 1 7.8457e+13 5.3388e+15 2901581

- ZINC2 1 1.6559e+14 5.4260e+15 2903495

- ZSMHC 1 7.7270e+14 6.0331e+15 2916039

> summary(stepwise.model)

complete.model= lm(

VALUE ~ ABL80 +

AGE1 +

B2000s +

B90s +

B40s50s + # What decade was the home built

B60s70s +

B80s +

B2010s +

BEDRMS +

ROOMS\_BEDS +

BEDRMS\_SQ +

BED\_PERSON +

ROOMS\_PERSON +

home\_age +

OTHERCOST +

PER +

PER\_SQ +

ROOMS +

ROOMS\_SQ +

struc1 +

struc2 +

struc3 +

struc4 +

struc5 +

TOTSAL +

TOTSAL\_PERSON +

UTILITY +

UTILITY\_SQ +

ZINC2 +

ZSMHC, data=thads

)