Stat 104 Fall 2015
Regression Project
Due 4pm December 8, 2015

Background

To determine what factors impact health care utilization spending, among elderly, as part of Medicare. This information can possibly help the government identify what sort of programs to implement to reduce future expenditures.

Data source: 2005 Medical Expenditures Panel Survey
Dataset ref: http://people.fas.harvard.edu/~mparzen/stat104/projectdataV1

Dataset Analysis:

The Dependent variable used in the dataset is Total Expense (totalexp).
The average expense based on the raw data provided is $8190.40 with a std dev of $11,746.62

There are 18 independent variables, of which several are categorical and nominal variables.

Quantitative Variables:
There are 6 quantitative variables which include -
Member demographics: age in years (age), years of education(educ), annual family income(income)
Health: bmi in kgs/cm2 (bmi), number of Dr. vists(dr_vists), number of hospital visits(hosp_vis)

Categorical variables:
There are 8 categorical variables related to
Member demographics: lives in msa(msa =1), male(male = 1)
Behavioral health: smoking(smoker = 1), limitation(phy_lim =1)
Chronic conditions: chronic heart disease(chd = 1), high cholesterol(high_col = 1),
diabetes (diabetes=1), high blood pressure(high_bp =1)

Nominal variables:
There are 4 nominal variables: marital status (marital), race (race_grp), senior health (srhealth), mental health (mntl_health)
Nominal variables have been converted into categorical variables having values (0/1) for the purposes of modeling:

Marital	mar_wid	mar_divc	mar_nvr
married	0	0	0
widowed	1	0	0
divorced	0	1	0
never mar	0	0	1

race_grp	race_blk	race_oth	race_hisp
White	0	0	0
Black	1	0	0
Other	0	1	0
hispanic	0	0	1

sr_health	sr_vrg	sr_good	sr_poor
excellent	0	0	0
very good	1	0	0
Good	0	1	0
Poor	0	0	1

mntl_health	mntl_vrg	mntl_good	mntl_poor
excellent	0	0	0
very good	1	0	0
good	0	1	0
Poor	0	0	1

Preliminary Diagnostics

A diagnostic of the variables to look for mutli-collinearity shows no strong correlation between variables. Based on this result we do not drop any variables.

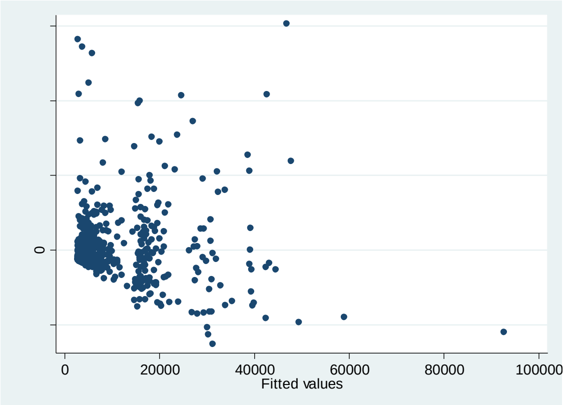
```
. cor totalexp age income educ
(obs=896)

             | totalexp      age   income     educ
-------------+------------------------------------
    totalexp |   1.0000
         age |   0.0763   1.0000
      income |  -0.0607  -0.1366   1.0000
        educ |  -0.0250  -0.0974   0.3821   1.0000

. cor totalexp bmi smoker phy_lim chd high_chol diabetes high_bp
(obs=896)

             | totalexp      bmi   smoker  phy_lim      chd high_c~l diabetes  high_bp
-------------+------------------------------------------------------------------------
    totalexp |   1.0000
         bmi |   0.0411   1.0000
      smoker |  -0.0079  -0.0704   1.0000
     phy_lim |   0.2223   0.1141  -0.0247   1.0000
         chd |   0.1928  -0.0007  -0.0419   0.1159   1.0000
   high_chol |   0.0579   0.1081  -0.0207  -0.0236   0.1483   1.0000
    diabetes |   0.1404   0.1623  -0.0233   0.0900   0.0590   0.1175   1.0000
     high_bp |   0.0538   0.1842  -0.0192   0.1090   0.1231   0.1422   0.1241   1.0000
```

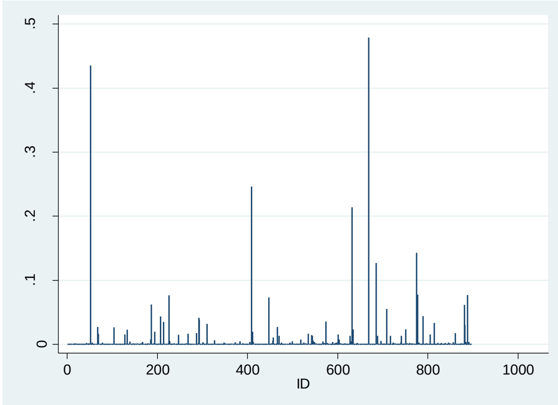
A visual inspection of relationships in the data reveals that there are transformations which would need to be applied to some of the variables. There are also outliers and influential points which need to be resolved.

To perform a third diagnostic for noise, we run a stepwise regression on the qualitative variables (age, educ, income and bmi) followed by a hetroskedaticity test. The results of the test show a P=0.000 which indicates there is hetroskedaticity (non-uniform noise) in the varibles.

```
. sw regress totalexp age educ income bmi dr_visits hosp_vis, pr(0.05)
                      begin with full model
p = 0.7933 >= 0.0500  removing income
p = 0.5925 >= 0.0500  removing bmi
p = 0.3864 >= 0.0500  removing educ
p = 0.2671 >= 0.0500  removing age

      Source |       SS           df       MS      Number of obs   =       896
-------------+----------------------------------   F(2, 893)       =    523.56
       Model |  6.6653e+10         2  3.3326e+10   Prob > F        =    0.0000
    Residual |  5.6842e+10       893  63652861.6   R-squared       =    0.5397
-------------+----------------------------------   Adj R-squared   =    0.5387
       Total |  1.2349e+11       895   137982985   Root MSE        =    7978.3

------------------------------------------------------------------------------
    totalexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hosp_vis |   11912.42   431.5609    27.60   0.000     11065.42    12759.41
   dr_visits |   231.7197   20.80119    11.14   0.000     190.8948    272.5446
       _cons |   2801.749   349.8589     8.01   0.000     2115.107     3488.39
------------------------------------------------------------------------------

. hettest

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

         chi2(1)      =   692.58
         Prob > chi2  =   0.0000
```

The rvfplot also shows hetroskedaticity and nonlinearity, which would require deleting influential data points





Using Cooks distance to identify influential points, we drop data poing where D_res> 0.1, we drop 6 data points.

Rerunning the regression and hettest, we do not find any difference with a P=0.000 on the hettest and as shown by the rvfplot in the results (as seen in figure1, graph below on the left).

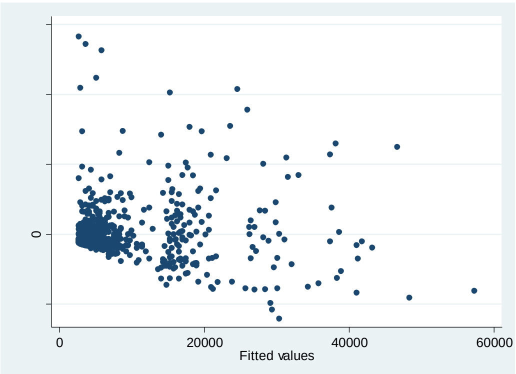


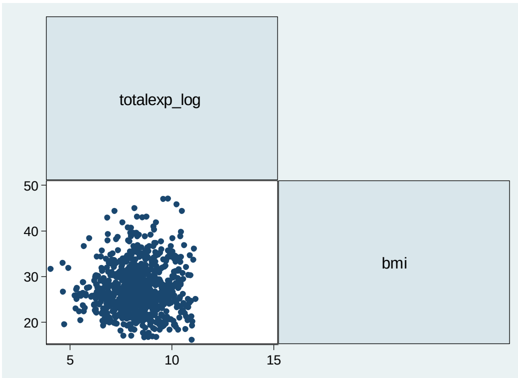
Figure 1: After Removing outliers

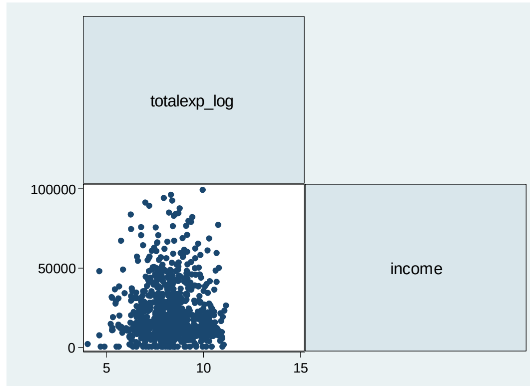
Figure 2: Using Log(Totalexp)

Trying a different option to log the response variable (totalexp), we get are able to resolve the issue of hetroskedaticity in the data as the P = 0.2808 (> 0.05) is now much higher.

However the rvfplot still shows non-linerarity (as seen in figure 2, graph above on the right). Running an ovtest shows a P=0.000, which indicates transformation is required in the X variables.

Further analysis of each of the quantitative variables reveals influential points which can be dropped.





Dropping extremely high bmi values: bmi>47 (5 points)
Dropping records where total expense is very low: totalexp<$115 (4points)
Dropping extremely high income values as average income is approx. $28,000 for the medicare population: income>$120,000 (7 points)

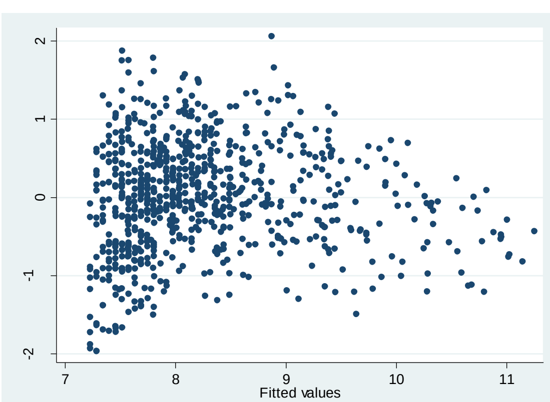
Running another regression and rvfplot does show some improvement, however, to increase the accuracy in the model we consider removing additional points, using Cooks distance (3 iterations).

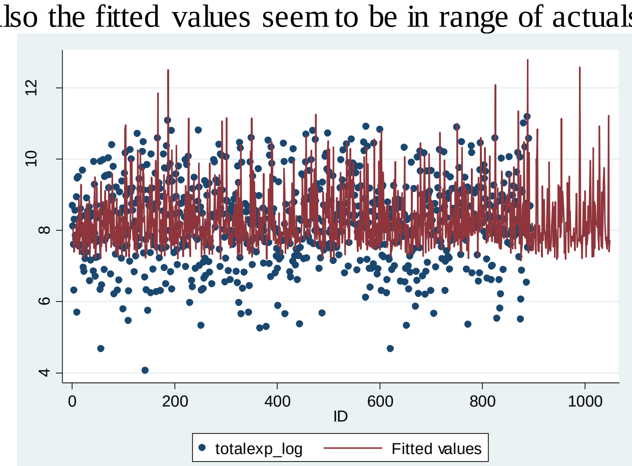
Generating residuals and dropping points where standardized residuals >2 helps reduce nonlinearity further.

Applying all the remaining dummy variables and rerunning regression gives a R=0.5331 with an Adjusted R=0.5293 and a Se = 0.69636 for totalexp_log

```
. . sw regress totalexp_log age educ income bmi_sq dr_visits hosp_vis msa male smoker phy_lim chd high_chol diabetes high_bp mar_wi
> d mar_divc mar_nvr race_blk race_oth race_hisp sr_vrg sr_good sr_poor mntl_vrg mntl_good mntl_poor, pr(0.1)
                      begin with full model
p = 0.9285 >= 0.1000  removing mar_divc
p = 0.9040 >= 0.1000  removing male
p = 0.8719 >= 0.1000  removing mar_nvr
p = 0.8446 >= 0.1000  removing mntl_vrg
p = 0.6808 >= 0.1000  removing sr_good
p = 0.6813 >= 0.1000  removing bmi_sq
p = 0.5902 >= 0.1000  removing smoker
p = 0.5986 >= 0.1000  removing sr_poor
p = 0.4433 >= 0.1000  removing mntl_good
p = 0.2751 >= 0.1000  removing msa
p = 0.2663 >= 0.1000  removing mntl_poor
p = 0.2055 >= 0.1000  removing high_bp
p = 0.1590 >= 0.1000  removing income
p = 0.1532 >= 0.1000  removing educ
p = 0.1309 >= 0.1000  removing race_oth
p = 0.1007 >= 0.1000  removing race_hisp
p = 0.1015 >= 0.1000  removing race_blk
p = 0.1157 >= 0.1000  removing sr_vrg
p = 0.1167 >= 0.1000  removing mar_wid

      Source |       SS           df       MS      Number of obs   =       859
-------------+----------------------------------   F(7, 851)       =    138.80
       Model |  612.077532         7  87.4396474   Prob > F        =    0.0000
    Residual |  536.086752       851  .62994918    R-squared       =    0.5331
-------------+----------------------------------   Adj R-squared   =    0.5293
       Total |  1148.16428       858  1.33818681   Root MSE        =    .79369

------------------------------------------------------------------------------
totalexp_log |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0120799    .004455     2.71   0.007     .0033359     .020824
    diabetes |   .2433446   .0697142     3.49   0.001     .1065127    .3801766
     phy_lim |     .28733   .0585325     4.91   0.000      .172445    .4022151
         chd |    .216087     .08187     2.64   0.008     .0553961    .3767778
   dr_visits |   .0492698   .0032378    15.22   0.000     .0429147    .0556248
    hosp_vis |    .903229   .0514573    17.55   0.000     .8022308    1.004227
   high_chol |   .2711478   .0561173     4.83   0.000     .1610034    .3812923
       _cons |   6.369955   .3328804    19.14   0.000     5.716592    7.023318
------------------------------------------------------------------------------

. . predict yhat_total_exp_log2
(option xb assumed; fitted values)

. . generate yhat_total_exp_invlog2 = exp(yhat_total_exp_log2)

. predict res_totalexp, r
(152 missing values generated)

. generate sres_4= res_totalexp/0.79369
(152 missing values generated)
```

Plotting an rvfplot shows residuals within 2. Also the fitted values seem to be in range of actuals.





The new mean and median information suggests the mean has not changed much, however variance has increased with the fitted data. The reason for the variance seems to be due some outliers in the fitted values.

```
. summarize yhat_total_exp_invlog2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
yhat_t~vlog2 |        997    7263.984     20142.97   1280.684   355901.3
```

On further analysis these points see to have significant differences, listed below
line # 990 has the highest predicted totalexp close to $300,000. This is due to the person having hospital visits=5
line # 954 and 1046 are similar profiles with high predicted totalexp close to $70,000. This is due to person having number of hospital or dr visits = 34. There is also a physical limitation indicator and high cholesterol indicator.

Based on the current model the following variable coefficients are included:
Age, Doctor visits, Hospital visits, diabetes, high cholesterol, chronic heart disease, physical limitation.

Major variables excluded are:
Age, income, educ, race, msa, bmi, martial, male, smoker, mntl_health, sr health

The equation for the response variable is given by:
Total_exp_log=6.37+0.903hosp_vis+0.057dr_visits+0.25diabetes+0.27high_chol+0.22chd+0.287phy_lim

The model has a variance of 0.794 which gives us a confidence interval for log total exp (-0.762, 2.35)

Conclusion:

Based on the above analysis we can determine key factors which influence the cost of medicare spending and what measures can be implemented to help save cost. Key influencers are the costs due to hospital visits and high number of dr visits. Members, especially those with severe chronic conditions, such as high cholesterol and diabetes, should be provided preventative care as this would reduce visits to the hospital in case of illness. It can also be inferred that demographics such as income, education and race are not that significant as compared to patients requiring treatment or being hospitalized.