# 705final

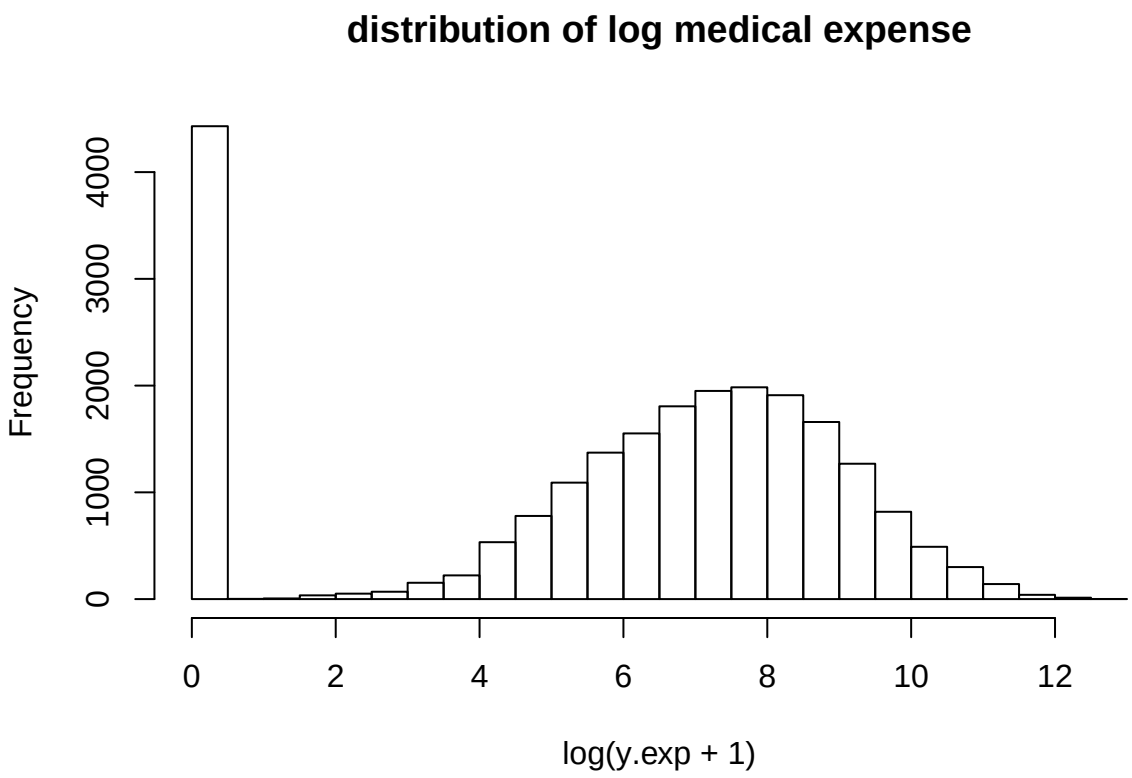
*ShanZhong*

*May 3, 2019*

```
## Loading required package: Matrix
## Loading required package: foreach
## Loaded glmnet 2.0-16
```

**first we load the data and remove id number**

**check the distribution of log expenditures**



## Create an algorithm to filter the data

## first check for missing value

## define

## filter data that have over 1000 missing values

```
##       list.name list.missing
##    1:     PANEL            0
##    2:  FCSZ1231          569
##    3:  FCRP1231           35
##    4:  RUSIZE31           27
##    5:  RUSIZE42            0
##   ---
## 1319:  SAQWT09F            0
## 1320:  DIABW09F            0
## 1321:    VARSTR            0
## 1322:    VARPSU            0
## 1323:      MSCD            0
```

## Detect factors in the data

## generate create factor and numeric data

## calculate pvalue for each of the factor variables

## take the average of estimate value and p-value

## calculate pvalue for each of the numeric variables

## choose a level of p-value that can filter unrelated variables

## utlize a lasso method to reduce parameters

```
##    user  system elapsed
##  147.27    0.01  147.31
```

```
##
## Call:  glmnet(x = as.matrix(dat.combine), y = as.numeric(y), family = "binomial")
##
##        Df       %Dev    Lambda
##  [1,]   0 -2.168e-13 1.669e-01
##  [2,]   2  3.800e-02 1.521e-01
##  [3,]   3  7.331e-02 1.386e-01
##  [4,]   4  1.081e-01 1.263e-01
##  [5,]   5  1.396e-01 1.151e-01
##  [6,]   6  1.683e-01 1.048e-01
##  [7,]   7  1.931e-01 9.553e-02
##  [8,]   7  2.142e-01 8.704e-02
##  [9,]   7  2.322e-01 7.931e-02
```

```
## [10,]   8  2.486e-01 7.226e-02
## [11,]   9  2.660e-01 6.584e-02
## [12,]  10  2.816e-01 5.999e-02
## [13,]  11  2.959e-01 5.466e-02
## [14,]  15  3.105e-01 4.981e-02
## [15,]  16  3.267e-01 4.538e-02
## [16,]  18  3.426e-01 4.135e-02
## [17,]  19  3.590e-01 3.768e-02
## [18,]  21  3.760e-01 3.433e-02
## [19,]  22  3.933e-01 3.128e-02
## [20,]  22  4.101e-01 2.850e-02
## [21,]  22  4.263e-01 2.597e-02
## [22,]  22  4.420e-01 2.366e-02
## [23,]  22  4.573e-01 2.156e-02
## [24,]  22  4.722e-01 1.965e-02
## [25,]  23  4.867e-01 1.790e-02
## [26,]  24  5.011e-01 1.631e-02
## [27,]  25  5.152e-01 1.486e-02
## [28,]  26  5.290e-01 1.354e-02
## [29,]  27  5.425e-01 1.234e-02
## [30,]  29  5.557e-01 1.124e-02
## [31,]  30  5.686e-01 1.024e-02
## [32,]  30  5.811e-01 9.333e-03
## [33,]  29  5.932e-01 8.504e-03
## [34,]  30  6.048e-01 7.749e-03
## [35,]  37  6.161e-01 7.060e-03
## [36,]  37  6.270e-01 6.433e-03
## [37,]  38  6.373e-01 5.862e-03
## [38,]  39  6.472e-01 5.341e-03
## [39,]  42  6.567e-01 4.866e-03
## [40,]  44  6.658e-01 4.434e-03
## [41,]  45  6.743e-01 4.040e-03
## [42,]  50  6.824e-01 3.681e-03
## [43,]  51  6.901e-01 3.354e-03
## [44,]  56  6.975e-01 3.056e-03
## [45,]  57  7.046e-01 2.785e-03
## [46,]  61  7.112e-01 2.537e-03
## [47,]  64  7.175e-01 2.312e-03
## [48,]  70  7.236e-01 2.107e-03
## [49,]  71  7.297e-01 1.919e-03
## [50,]  73  7.353e-01 1.749e-03
## [51,]  75  7.404e-01 1.594e-03
## [52,]  84  7.454e-01 1.452e-03
## [53,]  88  7.501e-01 1.323e-03
## [54,]  91  7.544e-01 1.205e-03
## [55,]  94  7.587e-01 1.098e-03
## [56,]  98  7.624e-01 1.001e-03
## [57,] 103  7.661e-01 9.119e-04
## [58,] 110  7.694e-01 8.309e-04
## [59,] 114  7.727e-01 7.570e-04
## [60,] 116  7.756e-01 6.898e-04
## [61,] 122  7.784e-01 6.285e-04
## [62,] 124  7.809e-01 5.727e-04
## [63,] 130  7.833e-01 5.218e-04
```

```
## [64,] 133  7.855e-01 4.754e-04
## [65,] 148  7.875e-01 4.332e-04
## [66,] 153  7.895e-01 3.947e-04
## [67,] 155  7.913e-01 3.597e-04
## [68,] 157  7.929e-01 3.277e-04
## [69,] 165  7.944e-01 2.986e-04
## [70,] 168  7.958e-01 2.721e-04
## [71,] 170  7.971e-01 2.479e-04
## [72,] 178  7.983e-01 2.259e-04
## [73,] 181  7.994e-01 2.058e-04
## [74,] 194  8.006e-01 1.875e-04
## [75,] 200  8.017e-01 1.709e-04
## [76,] 214  8.028e-01 1.557e-04
## [77,] 221  8.037e-01 1.419e-04
## [78,] 221  8.045e-01 1.293e-04
## [79,] 227  8.053e-01 1.178e-04
## [80,] 241  8.061e-01 1.073e-04
## [81,] 244  8.068e-01 9.778e-05
## [82,] 250  8.075e-01 8.909e-05
## [83,] 261  8.084e-01 8.118e-05
## [84,] 269  8.088e-01 7.396e-05
## [85,] 266  8.102e-01 6.739e-05
## [86,] 276  8.106e-01 6.141e-05
## [87,] 274  8.117e-01 5.595e-05
## [88,] 281  8.124e-01 5.098e-05
## [89,] 286  8.126e-01 4.645e-05
## [90,] 293  8.130e-01 4.233e-05
## [91,] 297  8.140e-01 3.856e-05
## [92,] 305  8.141e-01 3.514e-05
## [93,] 306  8.144e-01 3.202e-05
## [94,] 309  8.152e-01 2.917e-05
## [95,] 324  8.152e-01 2.658e-05
```

**from the lasso model, we find some interesting variables**

**Still we have some variables that are useless, we delete them**

**First we see the prediction performance(%Dev) based on number of variables choosed (DF)**

**then we fit a glm model with 80 parameters to choose significant variables**

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:  glm(formula = paste("y~", paste(variable, collapse = "+")), family = "binomial",
##     data = dat)
##
## Coefficients:
## (Intercept)     RUSIZE53     RUSIZE09     REFPRS09     ENDRFM53
##   3.204e-01    2.955e-02   -6.034e-02   -1.687e-03   -1.239e-02
```

```
##      AGE31X       EDUCYR     RFREL53X     BMINDX53        PCS42
##   3.359e-03    4.071e-02   -4.383e-03    1.087e-02   -4.223e-03
##    DDBDYS53     BUSNP09X     WCMPP09X     CHLDP09X     TOTTCH09
##   6.242e-03    2.453e-05    1.165e-04    1.531e-05    3.615e-04
##    OBTOTV09      OBDRV09     OBCHIR09     OBNURS09     OBETCH09
##  -2.895e-02    5.071e+00    3.364e+00    1.816e+00    5.757e-02
##    OBASST09      OPDRV09     AMCHIR09     AMASST09     AMTOTC09
##  -9.006e+00    5.296e+00    1.426e+00    1.349e+01    2.167e+00
##     ERTOT09      DVTOT09      DVGEN09     DVORTH09     DVOTCH09
##   4.553e+00    4.243e+00    6.416e-01   -4.601e+00    1.757e-02
##    VISTCH09     FAMWT09F     DIABW09F       VARSTR        PANEL
##   3.381e-02    2.932e-06    7.017e-05    1.583e-03   -9.572e-02
##    FCRP1231     RUCLAS31     RUCLAS53     RUCLAS09        MSA53
##   8.787e-02    3.433e-01    2.458e-01           NA   -2.316e-01
##       MSA09       RESP31       RESP42       RESP53       RESP09
##          NA   -3.469e-02   -8.702e-02    8.230e-02           NA
##     PROXY53      PROXY09     BEGRFY31     INSCOP42          SEX
##   5.775e-01           NA    5.702e-04   -2.003e-01    3.427e-01
##    RACETHNX     SPOUIN09        HIDEG     RTHLTH31     MNHLTH31
##   1.929e-01   -2.070e-01    2.871e-02    2.802e-02    1.219e-01
##      HIBPDX         MIDX       EMPHDX       CHOLDX       DIABDX
##  -9.258e-01    2.688e-01    4.698e-01   -5.133e-01   -6.061e-01
##    ADLHLP31     DENTCK53     BPCHEK53     EXRCIS53     PHYACT53
##  -3.099e-01    7.700e-02   -2.084e-01   -7.785e-02    1.285e-01
##    SEATBE53     ADILCR42     ADRTCR42     ADAPPT42     ADSMOK42
##   5.302e-02   -5.483e-02   -4.000e-01    1.444e-01    2.170e-01
##    ADDRBP42     ADREST42     ADINSA42     HAVEUS42     MDUNAB42
##  -3.266e-01    5.884e-03   -4.154e-02   -2.108e-01    2.122e-01
##    MDDLAY42     PMUNAB42     PMDLAY42      EMPST31     HRWGIM31
##   5.069e-01   -2.180e-01   -3.870e-01    2.628e-02   -1.923e-01
##    HRWGIM42     HRWGIM53     WAGIMP09     IRAIMP09     REFIMP09
##  -2.725e-01   -2.808e-01   -4.291e-02   -6.110e-02   -1.413e-02
##    CSHIMP09     SSIIMP09     OTHIMP09      OPAFE09      OPAMA09
##   1.275e-01    6.818e-02    3.231e-01   -5.847e-01    1.175e+00
##     STAJL09     PUBFE09X     PUBJU09X      PNGFE09      POUMA09
##  -1.147e+00   -3.253e-01    1.137e-01    5.141e-01    1.456e-01
##     HPDMY09      HPDJL09      HPOFE09     INSFE09X     INSDE09X
##  -1.575e+00    8.973e-01    3.690e-01   -9.574e-02   -1.657e-01
##     MCREV09      UNINS09     TRIST31X     MCRPD09X     OTPUBB42
##  -5.789e-02    1.341e-01    2.617e-01   -4.467e-02   -8.386e-01
##    INSAT42X       VARPSU
##   5.303e-02   -7.053e-02
##
## Degrees of Freedom: 22674 Total (i.e. Null);  22572 Residual
## Null Deviance:       22400
## Residual Deviance: 4482  AIC: 4688
```

## Then we choose variable manully

## Here is some interesting parameters I found

## family total income

FAMINC09 , WCMPP09X # child support CHLDP09X # uninsured UNINS09 # education HIDEG EDUCYR # wear eyeglasses WRGLAS42 # ASTHMA DIAGNOSIS ASTHDX # HIGH CHOLESTEROL DIAGNOSIS CHOLDX # HIGH BLOOD PRESSURE HIBPDX # DIABETES DIAGNOSIS DIABDX # PERCEIVED HEALTH STATUS RTHLTH42 RTHLTH53 # PERCEIVED mental HEALTH STATUS MNHLTH31 # race RACETHNX # real race: language the responsdent use when interviewed INTVLANG # Age AGE42X AGE09X # military servicers to adjust for age ACTDTY42 # TIME SNCE LST BLOOD PRES CHK BPCHEK53 # we use the SPOUID09 to represent marriage SPOUID09 # smoke ADSMOK42 # Weight DIABW09F PERWT09F BMINDX53 # DR CHECKED BLOOD PRESSURE ADDRBP42 # SEX SEX

## our model achieved a pred rate of 86%

```
## [1] 16
```

```
##
## Call:
## glm(formula = paste("y~", paste(variable.final, collapse = "+")),
##     family = "binomial", data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4946   0.1300   0.2821   0.4609   2.4878
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.792047   0.915034  -6.330 2.45e-10 ***
## WCMPP09X     0.110861   0.028377   3.907 9.36e-05 ***
## UNINS092     0.841660   0.048580  17.325  < 2e-16 ***
## EDUCYR       0.058210   0.007281   7.994 1.30e-15 ***
## HIBPDX1      2.056028   0.735505   2.795 0.005184 **
## HIBPDX2      1.199838   0.734332   1.634 0.102276
## MNHLTH311    0.698585   0.425380   1.642 0.100536
## MNHLTH312    0.772877   0.426011   1.814 0.069644 .
## MNHLTH313    0.987050   0.426451   2.315 0.020637 *
## MNHLTH314    1.355624   0.438471   3.092 0.001990 **
## MNHLTH315    1.589919   0.482805   3.293 0.000991 ***
## RACETHNX2   -0.169079   0.074195  -2.279 0.022677 *
## RACETHNX3    0.250546   0.100046   2.504 0.012269 *
## RACETHNX4    0.522642   0.075697   6.904 5.04e-12 ***
## INTVLANG2   -0.087240   0.078489  -1.111 0.266355
## INTVLANG3   -0.416198   0.129158  -3.222 0.001271 **
## INTVLANG91  -0.006587   0.231308  -0.028 0.977280
## AGE09X       0.007397   0.001642   4.505 6.65e-06 ***
## BPCHEK531    1.643713   0.098554  16.678  < 2e-16 ***
## BPCHEK532    0.205454   0.105021   1.956 0.050429 .
## BPCHEK533   -0.153173   0.117676  -1.302 0.193037
## BPCHEK534   -0.197666   0.147909  -1.336 0.181419
## BPCHEK535   -0.356650   0.139336  -2.560 0.010478 *
```

```
## BPCHEK536   -0.250019   0.143905  -1.737 0.082319 .
## SPOUID09995 -0.077254   0.046251  -1.670 0.094857 .
## ADSMOK422    0.256896   0.054656   4.700 2.60e-06 ***
## DIABW09F     0.146397   0.019075   7.675 1.66e-14 ***
## PERWT09F     0.163780   0.038426   4.262 2.02e-05 ***
## ADDRBP421    0.480189   0.143764   3.340 0.000837 ***
## ADDRBP422   -0.565298   0.146294  -3.864 0.000111 ***
## SEX2         0.636867   0.044406  14.342  < 2e-16 ***
## MSCD1        0.826862   0.155982   5.301 1.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22399  on 22674  degrees of freedom
## Residual deviance: 14346  on 22643  degrees of freedom
## AIC: 14410
##
## Number of Fisher Scoring iterations: 6
```

```
## [1] 0.8635061
```

# I do not investage much on exp size

```
##
## Call:
## lm(formula = paste("y.exp~", paste(variable.final, collapse = "+")),
##     data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3597 -0.9239  0.0132  0.9455  6.0926
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.5209797  0.7673616   5.892 3.89e-09 ***
## WCMPP09X     0.0849468  0.0121616   6.985 2.95e-12 ***
## UNINS092     0.7404139  0.0343445  21.558  < 2e-16 ***
## EDUCYR       0.0300569  0.0039467   7.616 2.75e-14 ***
## HIBPDX1      0.1607412  0.6608213   0.243 0.807819
## HIBPDX2     -0.1390335  0.6607742  -0.210 0.833350
## MNHLTH311   -0.4736839  0.3389336  -1.398 0.162259
## MNHLTH312   -0.4063380  0.3390559  -1.198 0.230761
## MNHLTH313   -0.2117527  0.3390386  -0.625 0.532262
## MNHLTH314    0.1504386  0.3407976   0.441 0.658906
## MNHLTH315    0.5968805  0.3478806   1.716 0.086222 .
## RACETHNX2   -0.1631866  0.0427434  -3.818 0.000135 ***
## RACETHNX3   -0.2474574  0.0557966  -4.435 9.26e-06 ***
## RACETHNX4    0.1565277  0.0416285   3.760 0.000170 ***
## INTVLANG2   -0.1359098  0.0509062  -2.670 0.007596 **
## INTVLANG3   -0.3182110  0.0875854  -3.633 0.000281 ***
## INTVLANG91  -0.2142246  0.1427964  -1.500 0.133577
## AGE09X       0.0151792  0.0007661  19.813  < 2e-16 ***
## BPCHEK531    0.3844666  0.0832797   4.617 3.93e-06 ***
```

```
## BPCHEK532   -0.4732359  0.0914028  -5.177 2.27e-07 ***
## BPCHEK533   -0.4870303  0.1091433  -4.462 8.16e-06 ***
## BPCHEK534   -0.4443636  0.1460667  -3.042 0.002352 **
## BPCHEK535   -0.3108232  0.1399185  -2.221 0.026332 *
## BPCHEK536   -0.1228210  0.1470341  -0.835 0.403547
## SPOUID09995  0.0477617  0.0230559   2.072 0.038320 *
## ADSMOK422    0.0618476  0.0301756   2.050 0.040419 *
## DIABW09F     0.0706118  0.0041838  16.877  < 2e-16 ***
## PERWT09F     0.0658048  0.0202414   3.251 0.001152 **
## ADDRBP421    0.1531442  0.0907750   1.687 0.091606 .
## ADDRBP422   -0.4739475  0.0978355  -4.844 1.28e-06 ***
## SEX2         0.3161047  0.0227799  13.876  < 2e-16 ***
## MSCD1        0.7182481  0.0397367  18.075  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.472 on 18213 degrees of freedom
## Multiple R-squared:  0.2943, Adjusted R-squared:  0.2931
## F-statistic: 245.1 on 31 and 18213 DF,  p-value: < 2.2e-16
```

## example of first 20

```
##  [1]   991.6905  1588.6455   823.1827   493.7461   330.5855  1666.6511
##  [7]  3000.2421  1315.7195  1627.5436  1290.2789  4278.8135 24037.5120
## [13]  3052.6292  3668.1913  6260.2542   437.7869  2544.1481   949.9147
## [19]  4648.4086  7296.7739
```

```
## [1] 2191.052
```

## generate a report for male and female

```
##            mean        sd        low        up
## 40male 1028.153  746.4256  -434.8415  2491.147
## 65male 2984.245 2693.2222 -2294.4701  8262.961
## 80male 5272.994 4159.8546 -2880.3209 13426.309
```

```
##              mean       sd        low        up
## 40female 1580.050 1242.116  -854.4971  4014.597
## 65female 3946.030 3077.172 -2085.2278  9977.288
## 80female 6983.977 4353.756 -1549.3847 15517.338
```

**our lasso model catched the variables we want, which is great!**