STAT230-HW2-Problem7

Yifan Zheng

2/18/2020

The main outcome of interest is "lnmeddol" which denotes the log of medical expenses. Use linear regression to investigate the relationship between the outcome and various important covariates. The solution of this problem is not unique, but do justify your choice of covariates and model, and interpret the results.

```
suppressPackageStartupMessages(library("car"))
suppressPackageStartupMessages(library(sampleSelection))
```

```
data("RandHIE")

# check several rows
head(RandHIE)
```

```
##   plan site coins tookphys year   zper black   income     xage female
## 1    3    1   100        0    1 125024     1 13748.76 42.87748      0
## 2    3    1   100        0    2 125024     1 13748.76 43.87748      0
## 3    3    1   100        0    3 125024     1 13748.76 44.87748      0
## 4    3    1   100        0    4 125024     1 13748.76 45.87748      0
## 5    3    1   100        0    5 125024     1 13748.76 46.87748      0
## 6    3    1   100        0    1 125025     1 13748.76 16.59138      0
##   educdec time  outpdol   drugdol suppdol mentdol inpdol    meddol totadm
## 1      12    1  0.00000  8.451119       0       0      0  8.451119      0
## 2      12    1 48.78706 13.288409       0       0      0 62.075470      0
## 3      12    1  0.00000  0.000000       0       0      0  0.000000      0
## 4      12    1  0.00000  0.000000       0       0      0  0.000000      0
## 5      12    1  0.00000  0.000000       0       0      0  0.000000      0
## 6      12    1  0.00000  0.000000       0       0      0  0.000000      0
##   inpmis mentvis mdvis notmdvis num  mhi    disea physlm ghindx mdeoff
## 1      0       0     0        0   4 95.0 13.73189      0     NA   1000
## 2      0       0     2        0   4 95.0 13.73189      0     NA   1000
## 3      0       0     0        0   4 95.0 13.73189      0     NA   1000
## 4      0       0     0        0   4 95.0 13.73189      0     NA   1000
## 5      0       0     0        0   4 95.0 13.73189      0     NA   1000
## 6      0       0     0        0   4 93.8 13.73189      0     NA   1000
##   pioff child fchild     lfam      lpi idp logc fmde hlthg hlthf hlthp
## 1  1000     0      0 1.386294 6.907755   1    0    0     1     0     0
## 2  1000     0      0 1.386294 6.907755   1    0    0     1     0     0
## 3  1000     0      0 1.386294 6.907755   1    0    0     1     0     0
## 4  1000     0      0 1.386294 6.907755   1    0    0     1     0     0
## 5  1000     0      0 1.386294 6.907755   1    0    0     1     0     0
## 6  1000     1      0 1.386294 6.907755   1    0    0     0     0     0
##    xghindx     linc     lnum lnmeddol binexp
## 1 65.20780 9.528776 1.386294 2.134299      1
## 2 65.20780 9.528776 1.386294 4.128351      1
## 3 65.20780 9.528776 1.386294       NA      0
## 4 65.20780 9.528776 1.386294       NA      0
## 5 65.20780 9.528776 1.386294       NA      0
## 6 76.34753 9.528776 1.386294       NA      0
```

```r
# check missing values
sum(is.na(RandHIE))
```

```
## [1] 9690
```

```r
# delete missing values
Rand4lr <- na.omit(RandHIE)

# check if use all covariates for linear regression
lm_all <- lm(lnmeddol ~ ., data=Rand4lr)
summary(lm_all)
```

```
## 
## Call:
## lm(formula = lnmeddol ~ ., data = Rand4lr)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -13.5190  -0.4339   0.1717   0.5576   3.9578 
## 
## Coefficients: (3 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.719e+00  4.368e-01   3.936 8.32e-05 ***
## plan        -4.141e-03  3.130e-03  -1.323 0.185910    
## site        -3.931e-01  4.614e-01  -0.852 0.394230    
## coins       -2.357e-03  2.246e-03  -1.049 0.294059    
## tookphys     1.667e-02  1.851e-02   0.900 0.368078    
## year         4.885e-03  7.321e-03   0.667 0.504665    
## zper         3.718e-06  4.600e-06   0.808 0.418892    
## black       -5.487e-02  2.990e-02  -1.835 0.066534 .  
## income       7.318e-06  3.267e-06   2.240 0.025113 *  
## xage         2.917e-03  9.626e-04   3.030 0.002449 ** 
## female       1.409e-01  2.301e-02   6.125 9.39e-10 ***
## educdec     -3.821e-03  3.391e-03  -1.127 0.259795    
## time         1.898e+00  3.954e-01   4.800 1.61e-06 ***
## outpdol      1.640e-01  7.185e-02   2.282 0.022501 *  
## drugdol      1.631e-01  7.185e-02   2.270 0.023218 *  
## suppdol      1.684e-01  7.185e-02   2.344 0.019118 *  
## mentdol     -2.130e-04  3.504e-04  -0.608 0.543256    
## inpdol       1.605e-01  7.185e-02   2.235 0.025466 *  
## meddol      -1.604e-01  7.185e-02  -2.233 0.025596 *  
## totadm       1.297e+00  2.498e-02  51.911  < 2e-16 ***
## inpmis       2.521e-01  1.197e-01   2.106 0.035241 *  
## mentvis      9.362e-03  5.979e-03   1.566 0.117458    
## mdvis        5.507e-02  2.631e-03  20.932  < 2e-16 ***
## notmdvis     1.042e-02  2.255e-03   4.619 3.89e-06 ***
## num         -8.044e-02  1.479e-02  -5.440 5.42e-08 ***
## mhi         -3.746e-04  7.757e-04  -0.483 0.629121    
## disea        2.225e-03  1.432e-03   1.553 0.120398    
## physlm       8.347e-03  2.738e-02   0.305 0.760514    
## ghindx      -2.883e-03  6.810e-04  -4.233 2.32e-05 ***
## mdeoff      -1.608e-04  1.032e-04  -1.558 0.119163    
## pioff        9.538e-05  6.873e-05   1.388 0.165285    
## child       -2.047e-01  3.746e-02  -5.464 4.76e-08 ***
## fchild      -1.352e-01  3.651e-02  -3.702 0.000215 ***
## lfam         1.868e-01  5.049e-02   3.700 0.000216 ***
## lpi         -7.652e-03  6.920e-03  -1.106 0.268863
## idp         -1.571e-01  2.041e-01  -0.769 0.441613
## logc         3.912e-02  8.121e-02   0.482 0.629977
## fmde        -9.634e-03  3.305e-02  -0.291 0.770693
## hlthg        5.472e-02  2.017e-02   2.713 0.006686 **
## hlthf        3.133e-02  3.575e-02   0.877 0.380750
## hlthp       -1.319e-01  7.042e-02  -1.873 0.061068 .
## xghindx             NA         NA      NA       NA
## linc         1.146e-03  1.071e-02   0.107 0.914810
## lnum                NA         NA      NA       NA
## binexp              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.916 on 11457 degrees of freedom
## Multiple R-squared:  0.6278, Adjusted R-squared:  0.6265
## F-statistic: 471.4 on 41 and 11457 DF,  p-value: < 2.2e-16
```

F-statistic = 471.4 and p-value is very small (less than 0.05). It means with all covariates the linear model is better than the intercept-only model.

R-squared is 62.78%. It means the model explains 62.78% variation of response.

Now pick those covariates with small p-values (<0.05) of t-tests for a new linear model

```
cov_1 <- lnmeddol ~ income + xage + female + time + outpdol + drugdol +
    suppdol  + inpdol + meddol + totadm + inpmis +
    mdvis + notmdvis + num + ghindx + child + fchild + lfam + hlthg

lm_1 <- lm(cov_1, data=Rand4lr)
summary(lm_1)
```

```
##
## Call:
## lm(formula = cov_1, data = Rand4lr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6670  -0.4416   0.1822   0.5689   3.9463
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.618e+00  4.029e-01   4.015 5.97e-05 ***
## income       7.813e-06  2.340e-06   3.338 0.000846 ***
## xage         2.728e-03  9.030e-04   3.021 0.002522 **
## female       1.406e-01  2.211e-02   6.360 2.10e-10 ***
## time         1.886e+00  3.968e-01   4.753 2.03e-06 ***
## outpdol      1.712e-01  7.211e-02   2.374 0.017616 *
## drugdol      1.704e-01  7.211e-02   2.364 0.018118 *
## suppdol      1.758e-01  7.210e-02   2.438 0.014796 *
## inpdol       1.677e-01  7.211e-02   2.326 0.020048 *
## meddol      -1.676e-01  7.211e-02  -2.324 0.020152 *  
## totadm       1.293e+00  2.500e-02  51.724  < 2e-16 ***
## inpmis       1.833e-01  1.198e-01   1.530 0.126149    
## mdvis        5.654e-02  2.630e-03  21.496  < 2e-16 ***
## notmdvis     1.073e-02  2.244e-03   4.781 1.77e-06 ***
## num         -7.776e-02  1.451e-02  -5.361 8.46e-08 ***
## ghindx      -2.966e-03  5.716e-04  -5.189 2.15e-07 ***
## child       -2.090e-01  3.568e-02  -5.858 4.81e-09 ***
## fchild      -1.341e-01  3.602e-02  -3.722 0.000199 ***
## lfam         1.623e-01  4.943e-02   3.283 0.001029 ** 
## hlthg        5.682e-02  1.846e-02   3.077 0.002093 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9201 on 11479 degrees of freedom
## Multiple R-squared:  0.6237, Adjusted R-squared:  0.6231 
## F-statistic:  1002 on 19 and 11479 DF,  p-value: < 2.2e-16
```

In the updated model with less covariates, “inpmis” is not significant here (with a high p value for t test), and R-squared doesn’t change compared with former model.

Here check if I could delete “inpmis” from covariates, use F test to test the overall covariates with/without “inpmis”

```r
linearHypothesis(lm_1, "inpmis=0")
```

```
## Linear hypothesis test
## 
## Hypothesis:
## inpmis = 0
## 
## Model 1: restricted model
## Model 2: lnmeddol ~ income + xage + female + time + outpdol + drugdol + 
##     suppdol + inpdol + meddol + totadm + inpmis + mdvis + notmdvis + 
##     num + ghindx + child + fchild + lfam + hlthg
## 
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1  11480 9719.3                           
## 2  11479 9717.3  1    1.9805 2.3396 0.1261
```

Here P value is larger than 0.05, hence we cannot reject null hypothesis.

Then delete “inpmis” and update the model

```r
cov_2 <- lnmeddol ~ income + xage + female + time + outpdol + drugdol +
    suppdol  + inpdol + meddol + totadm +
    mdvis + notmdvis + num + ghindx + child + fchild + lfam + hlthg

lm_2 <- lm(cov_2, data=Rand4lr)
summary(lm_2)
```

```
## 
## Call:
```

```
## lm(formula = cov_2, data = Rand4lr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6806  -0.4400   0.1812   0.5695   3.9347
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.629e+00  4.028e-01   4.043 5.30e-05 ***
## income       7.725e-06  2.340e-06   3.301 0.000965 ***
## xage         2.704e-03  9.029e-04   2.994 0.002757 **
## female       1.411e-01  2.211e-02   6.382 1.82e-10 ***
## time         1.877e+00  3.967e-01   4.730 2.27e-06 ***
## outpdol      1.714e-01  7.212e-02   2.377 0.017456 *
## drugdol      1.707e-01  7.211e-02   2.367 0.017955 *
## suppdol      1.760e-01  7.211e-02   2.441 0.014661 *
## inpdol       1.680e-01  7.211e-02   2.329 0.019868 *
## meddol      -1.678e-01  7.211e-02  -2.327 0.019970 *
## totadm       1.300e+00  2.454e-02  53.000  < 2e-16 ***
## mdvis        5.649e-02  2.630e-03  21.477  < 2e-16 ***
## notmdvis     1.072e-02  2.244e-03   4.776 1.81e-06 ***
## num         -7.744e-02  1.451e-02  -5.339 9.54e-08 ***
## ghindx      -2.969e-03  5.716e-04  -5.193 2.10e-07 ***
## child       -2.103e-01  3.568e-02  -5.894 3.87e-09 ***
## fchild      -1.345e-01  3.602e-02  -3.733 0.000190 ***
## lfam         1.618e-01  4.943e-02   3.273 0.001066 **
## hlthg        5.639e-02  1.846e-02   3.054 0.002260 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9201 on 11480 degrees of freedom
## Multiple R-squared:  0.6237, Adjusted R-squared:  0.6231
## F-statistic:  1057 on 18 and 11480 DF,  p-value: < 2.2e-16
```

Now all covariates are significant, and the linear model is:

```
lm_2
```

```
##
## Call:
## lm(formula = cov_2, data = Rand4lr)
##
## Coefficients:
## (Intercept)       income         xage       female         time
##   1.629e+00    7.725e-06    2.704e-03    1.411e-01    1.877e+00
##     outpdol      drugdol      suppdol       inpdol       meddol
##   1.714e-01    1.707e-01    1.760e-01    1.680e-01   -1.678e-01
##      totadm        mdvis     notmdvis          num       ghindx
##   1.300e+00    5.649e-02    1.072e-02   -7.744e-02   -2.969e-03
##       child       fchild         lfam        hlthg
##  -2.103e-01   -1.345e-01    1.618e-01    5.639e-02
```