modified

WangYIng

2020/12/22

```
#read and preview data
df=read.csv('C:/Users/mac/Desktop/dataanalysis-R/class/final/csv/incomedata.csv',header = TRUE,
sep=',',na.strings = 'NA')
head(df)
```

```
##   city year treat urban    income
## 1 晋中 2013     1     1 14312.96
## 2 运城 2013     1     1  2803.95
## 3 临汾 2013     1     1  8742.28
## 4 金华 2013     1     1 11679.44
## 5 衢州 2013     1     1 -6573.55
## 6 萍乡 2013     1     1  8448.00
```

```
#drop the data whose income is less than zero
df=subset(df,income>0)
#add variable "post" which represent its high speed railway is constructed after 2014
df$post=ifelse(df$year>2014,1,0)
sapply(df,class)
```

```
##        city        year       treat       urban      income        post
## "character"   "integer"   "integer"   "integer" "character"   "numeric"
```

```
#convert the types of data to numeric
df=as.data.frame(lapply(df[,2:6],as.numeric))
#check the outcome
sapply(df,class)
```

```
##      year     treat     urban    income      post
## "numeric" "numeric" "numeric" "numeric" "numeric"
```

```
#DID model
df$'income_indicator'=df$income/1000
reg.urban=lm(income_indicator~post*treat,data=subset(df,urban==1))
reg.urban
```

```
##
## Call:
## lm(formula = income_indicator ~ post * treat, data = subset(df,
##     urban == 1))
##
## Coefficients:
## (Intercept)         post        treat   post:treat
##       8.846        4.900        2.345        2.157
```

```
reg.rural=lm(income_indicator~post*treat,data=subset(df,urban==0))
reg.rural
```

```
##
## Call:
## lm(formula = income_indicator ~ post * treat, data = subset(df,
##     urban == 0))
##
## Coefficients:
## (Intercept)         post        treat   post:treat
##       6.058        9.476        1.581       -4.830
```

```
#the result shows that the construction of HSR had supportive effect on the income of urban peo
ple
#but had negative effect on that of rural people
```

```
#calculate the average treat effect(ATE) of the opening of HSR on the income gap between urban
 and rural people
income_gap=reg.urban$coefficients[4]-reg.rural$coefficients[4]
income_gap
```

```
## post:treat
##   6.987684
```

```
#positive means that the construction of HSR did enhance the income gap
```

```
#make a table to check the DID model
dif.urban=rbind(c(reg.urban$coefficients[1],reg.urban$coefficients[1]+reg.urban$coefficients[2
],reg.urban$coefficients[2]),c(reg.urban$coefficients[1]+reg.urban$coefficients[3],reg.urban$co
efficients[1]+reg.urban$coefficients[3]+reg.urban$coefficients[2]+reg.urban$coefficients[4],re
g.urban$coefficients[2]+reg.urban$coefficients[4]),c(reg.urban$coefficients[3],reg.urban$coeffi
cients[3]+reg.urban$coefficients[4],reg.urban$coefficients[4]))

rownames(dif.urban)=c("Control","Treatment","Difference")
colnames(dif.urban)=c("Pre","Post","Difference")
dif.urban
```

```
##                  Pre      Post Difference
## Control     8.846411 13.746605   4.900194
## Treatment  11.191388 18.249042   7.057654
## Difference  2.344977  4.502437   2.157460
```

```
#the table shows that the cross-section estimate for the post-period is 4.5, which is positive
 and higher than that for the pre-period(2.3)
#meaning that the construction of HSR had a positive impact on the urban people's income
```

```
#make a table to check the DID model
dif.rural=rbind(c(reg.rural$coefficients[1],reg.rural$coefficients[1]+reg.rural$coefficients[2
],reg.rural$coefficients[2]),c(reg.rural$coefficients[1]+reg.rural$coefficients[3],reg.rural$co
efficients[1]+reg.rural$coefficients[3]+reg.rural$coefficients[2]+reg.rural$coefficients[4],re
g.rural$coefficients[2]+reg.rural$coefficients[4]),c(reg.rural$coefficients[3],reg.rural$coeffi
cients[3]+reg.rural$coefficients[4],reg.rural$coefficients[4]))

rownames(dif.rural)=c("Control","Treatment","Difference")
colnames(dif.rural)=c("Pre","Post","Difference")
dif.rural
```

```
##                 Pre      Post Difference
## Control    6.058475 15.534197   9.475723
## Treatment  7.639698 12.285197   4.645499
## Difference 1.581223 -3.249001  -4.830224
```

```
#the table shows that the cross-section estimate for the post-period is -3.2, which is negative
and less than that for the pre-period(1.6)
#meaning that the construction of HSR had a negative impact on the income of rural people
```

```
#DDD model used to recheck ATE and conclusions above
reg.triple=lm(income_indicator~post*treat*urban,data=df)
reg.triple
```

```
##
## Call:
## lm(formula = income_indicator ~ post * treat * urban, data = df)
##
## Coefficients:
##      (Intercept)              post             treat             urban
##           6.0585            9.4757            1.5812            2.7879
##       post:treat        post:urban       treat:urban  post:treat:urban
##          -4.8302           -4.5755            0.7638            6.9877
```

```
stargazer(reg.urban, reg.rural, reg.triple, type="text",title = "Influence Analysis",covariate.
labels = c('POST','HSR','urban','HSR x Post','Post x Urban','HSR x Urban','HSR x Post x Urban'
), dep.var.labels= c("urban  VS  rural  VS triple"), omit.stat = c("ser",'rsq',"adj.rsq"))
```

```
##
## Influence Analysis
## ========================================================================================
##                                           Dependent variable:
##                     --------------------------------------------------------------------
##                                        urban VS rural VS triple
##                              (1)                    (2)                    (3)
## ----------------------------------------------------------------------------------------
## POST                      4.900**                9.476***               9.476***
##                           (2.388)                (2.986)                (2.710)
##
## HSR                        2.345                  1.581                  1.581
##                           (3.245)                (4.736)                (4.299)
##
## urban                                                                    2.788
##                                                                         (2.886)
##
## HSR x Post                 2.157                  -4.830                 -4.830
##                           (4.260)                (6.299)                (5.718)
##
## Post x Urban                                                             -4.576
##                                                                         (3.778)
##
## HSR x Urban                                                              0.764
##                                                                         (5.593)
##
## HSR x Post x Urban                                                       6.988
##                                                                         (7.399)
##
## Constant                  8.846***               6.058***               6.058***
##                           (1.827)                (2.277)                (2.067)
##
## ----------------------------------------------------------------------------------------
## Observations                492                    409                    901
## F Statistic       3.691** (df = 3; 488) 3.652** (df = 3; 405) 3.336*** (df = 7; 893)
## ========================================================================================
## Note:                                                      *p<0.1; **p<0.05; ***p<0.01
```

```
#according to the outcomes above, we can see that the opening of HSR had a positive effect on the income of urban residents, but a negative effect on the income of rural residents, thus exacerbating the urban-rural income gap
#We speculate that the negative impact of HSR on rural areas was caused by the loss of rural labor force, considering that the opening of HSR will intensify the amount of labor force migrate to cities
#then We'll try to verify this assumption
```

```
#assumption verifying
#relevance test
#read and preview data
data=read.csv('C:/Users/mac/Desktop/dataanalysis-R/class/final/csv/allcity.csv')
head(data,n=10)
```

```
##      city year      income     rev exp_gen exp_sci exp_edu  popu labor
## 1  石家庄 2013 20109.05346 1803141 2329600   45086  520170 246.9  55.8
## 2    唐山 2013  11632.2682 1995006 3044884   65704  546036 322.8  62.4
## 3  秦皇岛 2013 28310.29076  855537 1283636   10817  187008  90.9  24.1
## 4    邯郸 2013 9451.816381  851184 1239402   14772  261214 148.5  29.1
## 5    邢台 2013 11483.18513  366097  682631    3325  145951  91.2  16.9
## 6    保定 2013  16633.5564  701322 1013202   12343  143223 108.2  33.0
## 7  张家口 2013 16330.63819  133081  322020    2476   70043  90.0  18.0
## 8    承德 2013 16114.16965  409986  842145    3762  150218  58.7  11.1
## 9    沧州 2013 18303.79681  595128 1092589    8779  215631  53.9  17.0
## 10   廊坊 2013 16054.24941  648408  956480   11035  183614  81.3  18.9
##    inds_firm
## 1        245
## 2        602
## 3        233
## 4        173
## 5         80
## 6        194
## 7        133
## 8         89
## 9        167
## 10       211
```

```
#select the data, add variables and convert data types to numeric
newdata=data[,3:10]
sapply(newdata,class)
```

```
##      income         rev     exp_gen     exp_sci     exp_edu        popu
## "character"   "integer"   "numeric"   "integer" "character"   "numeric"
##       labor   inds_firm
##   "numeric"   "integer"
```

```
newdata=as.data.frame(lapply(newdata,as.numeric))
```

```
## Warning in lapply(newdata, as.numeric): 强制改变过程中产生了NA

## Warning in lapply(newdata, as.numeric): 强制改变过程中产生了NA
```

```
sapply(newdata,class)
```

```
##    income       rev   exp_gen   exp_sci   exp_edu      popu     labor inds_firm
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
```

```
newdata=na.omit(newdata)
head(newdata)
```

```
##      income     rev exp_gen exp_sci exp_edu  popu labor inds_firm
## 1 20109.053 1803141 2329600   45086  520170 246.9  55.8       245
## 2 11632.268 1995006 3044884   65704  546036 322.8  62.4       602
## 3 28310.291  855537 1283636   10817  187008  90.9  24.1       233
## 4  9451.816  851184 1239402   14772  261214 148.5  29.1       173
## 5 11483.185  366097  682631    3325  145951  91.2  16.9        80
## 6 16633.556  701322 1013202   12343  143223 108.2  33.0       194
```

```
#correlation analysis for the putative income per capita, fiscal revenue, three kinds of fiscal
expense,population, labor force and amount of local enterprise
cor(newdata)
```

```
##               income       rev   exp_gen    exp_sci   exp_edu      popu
## income    1.00000000 0.1561888 0.1695489 0.08762185 0.1616800 0.1916599
## rev       0.15618884 1.0000000 0.9624189 0.87189734 0.9526988 0.6999813
## exp_gen   0.16954894 0.9624189 1.0000000 0.89363937 0.9378381 0.6796498
## exp_sci   0.08762185 0.8718973 0.8936394 1.00000000 0.8142512 0.4583796
## exp_edu   0.16168004 0.9526988 0.9378381 0.81425124 1.0000000 0.7782290
## popu      0.19165993 0.6999813 0.6796498 0.45837958 0.7782290 1.0000000
## labor     0.24937144 0.7568898 0.7615744 0.67140320 0.7573079 0.5795051
## inds_firm 0.10485442 0.7838682 0.7362816 0.65668682 0.8265520 0.6211472
##               labor inds_firm
## income    0.2493714 0.1048544
## rev       0.7568898 0.7838682
## exp_gen   0.7615744 0.7362816
## exp_sci   0.6714032 0.6566868
## exp_edu   0.7573079 0.8265520
## popu      0.5795051 0.6211472
## labor     1.0000000 0.6047323
## inds_firm 0.6047323 1.0000000
```

```
corr.test(newdata, use="complete")
```

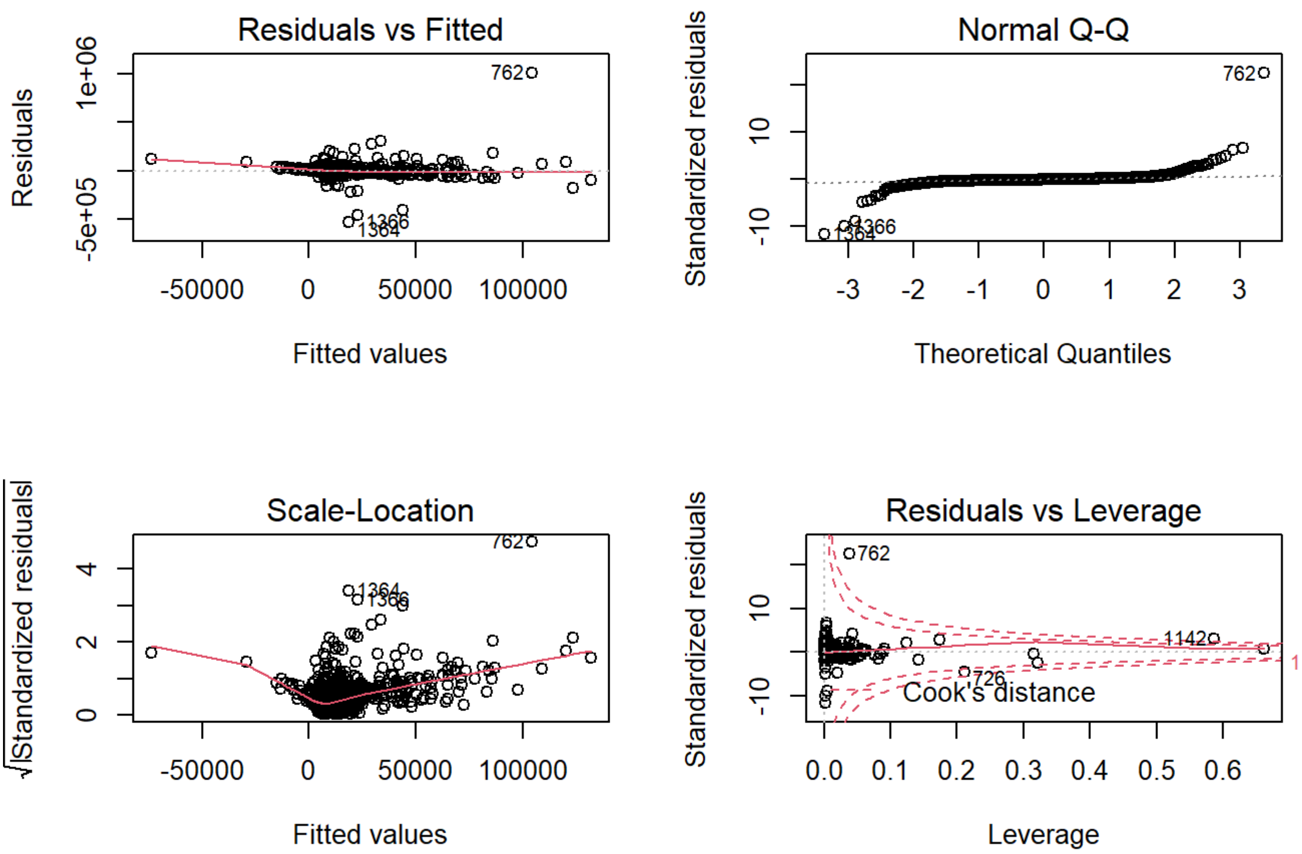
```
## Call:corr.test(x = newdata, use = "complete")
## Correlation matrix
##           income  rev exp_gen exp_sci exp_edu popu labor inds_firm
## income      1.00 0.16    0.17    0.09    0.16 0.19  0.25      0.10
## rev         0.16 1.00    0.96    0.87    0.95 0.70  0.76      0.78
## exp_gen     0.17 0.96    1.00    0.89    0.94 0.68  0.76      0.74
## exp_sci     0.09 0.87    0.89    1.00    0.81 0.46  0.67      0.66
## exp_edu     0.16 0.95    0.94    0.81    1.00 0.78  0.76      0.83
## popu        0.19 0.70    0.68    0.46    0.78 1.00  0.58      0.62
## labor       0.25 0.76    0.76    0.67    0.76 0.58  1.00      0.60
## inds_firm   0.10 0.78    0.74    0.66    0.83 0.62  0.60      1.00
## Sample Size
## [1] 1312
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##           income rev exp_gen exp_sci exp_edu popu labor inds_firm
## income         0   0       0       0       0    0     0         0
## rev            0   0       0       0       0    0     0         0
## exp_gen        0   0       0       0       0    0     0         0
## exp_sci        0   0       0       0       0    0     0         0
## exp_edu        0   0       0       0       0    0     0         0
## popu           0   0       0       0       0    0     0         0
## labor          0   0       0       0       0    0     0         0
## inds_firm      0   0       0       0       0    0     0         0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

```
#linear regression
myfit=lm(income~rev+exp_gen+exp_sci+exp_edu+popu+labor+inds_firm,newdata)
summary(myfit)
```

```
##
## Call:
## lm(formula = income ~ rev + exp_gen + exp_sci + exp_edu + popu +
##     labor + inds_firm, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -528531   -7159   -1642    5381 1005703
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.608e+03  2.100e+03   1.718 0.086053 .
## rev         -1.011e-03  2.536e-03  -0.399 0.690197
## exp_gen      6.616e-03  2.225e-03   2.974 0.002998 **
## exp_sci     -7.259e-02  1.982e-02  -3.663 0.000259 ***
## exp_edu     -2.642e-02  1.642e-02  -1.609 0.107856
## popu         4.284e+01  2.052e+01   2.088 0.036995 *
## labor        3.307e-02  4.719e-03   7.009 3.84e-12 ***
## inds_firm   -1.854e+00  2.622e+00  -0.707 0.479682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45570 on 1304 degrees of freedom
## Multiple R-squared:  0.08801,    Adjusted R-squared:  0.08311
## F-statistic: 17.98 on 7 and 1304 DF,  p-value: < 2.2e-16
```

```
#it can be seen that the linear relationships between general fiscal expenditure, scientific fi
scal expenditure, population, labor force and income per capita are significant
```

```
#regression diagnosis
par(mfrow=c(2,2))
plot(myfit)
```




```
#it does not satisfy the assumption of normal distribution and has some outliers
```

```
#outlier testing
sqrt(vif(myfit))
```

```
##      rev  exp_gen  exp_sci  exp_edu     popu    labor inds_firm
## 4.491978 4.468214 2.570905 4.496519 1.835473 1.574563  1.843837
```

```
outlierTest(myfit)
```

```
##        rstudent unadjusted p-value Bonferroni p
## 762   28.751809        3.4724e-141  4.5558e-138
## 1364 -12.252841         9.5274e-33   1.2500e-29
## 1366 -10.395414         2.2640e-24   2.9703e-21
## 1019  -9.140202         2.3228e-19   3.0475e-16
## 804    6.845897         1.1672e-11   1.5314e-08
## 1307   6.232047         6.2040e-10   8.1396e-07
## 1008   4.974125         7.4288e-07   9.7466e-04
## 868   -4.955490         8.1629e-07   1.0710e-03
## 214   -4.662868         3.4388e-06   4.5118e-03
## 1308   4.491582         7.6956e-06   1.0097e-02
```

```
newdata1=newdata[c(-762,-1364,-1366,-1019,-804,-1307,-1008,-868,-214,-1308),]
#and 10 outliers havd been found and dropped
```

```
#multicollinearity test
vif(myfit)
```

```
##       rev   exp_gen   exp_sci   exp_edu      popu     labor inds_firm
## 20.177867 19.964934  6.609555 20.218687  3.368963  2.479250  3.399734
```

```
sqrt(vif(myfit))>2
```

```
##       rev   exp_gen   exp_sci   exp_edu      popu     labor inds_firm
##      TRUE      TRUE      TRUE      TRUE     FALSE     FALSE     FALSE
```

```
#it shows that four kinds of variables fiscal revenue and expenditure are multicollinear, so ma
ke Principal Component Analysis to three variables for expenditure, given that the regression c
oefficient of revenue is not significant
```

```
#Principal Component Analysis and use the first principal component to replace the overall fisc
al expenditure
pca=princomp(newdata[3:5],cor=T)
summary(pca,loadings=T)
```

```
## Importance of components:
##                           Comp.1     Comp.2     Comp.3
## Standard deviation     1.6627673 0.43604246 0.21230163
## Proportion of Variance 0.9215983 0.06337767 0.01502399
## Cumulative Proportion  0.9215983 0.98497601 1.00000000
##
## Loadings:
##         Comp.1 Comp.2 Comp.3
## exp_gen  0.592  0.138  0.794
## exp_sci  0.565 -0.774 -0.286
## exp_edu  0.575  0.618 -0.536
```

```
pca_data=as.data.frame(predict(pca))
write.csv(pca_data,"pcadata.csv")
#save the outcomes of PCA into csv file and then restore standardized data to the original form in Excel
```

```
#format adjusting
sapply(data,class)
```

```
##        city        year      income         rev     exp_gen     exp_sci
## "character"   "integer" "character"   "integer"   "numeric"   "integer"
##     exp_edu        popu       labor   inds_firm
## "character"   "numeric"   "numeric"   "integer"
```

```
temp=as.data.frame(lapply(data[,2:10],as.numeric))
```

```
## Warning in lapply(data[, 2:10], as.numeric): 强制改变过程中产生了NA

## Warning in lapply(data[, 2:10], as.numeric): 强制改变过程中产生了NA
```

```
sapply(temp,class)
```

```
##      year    income       rev   exp_gen   exp_sci   exp_edu      popu     labor
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
## inds_firm
## "numeric"
```

```
temp=cbind(data[,1],temp)
temp=na.omit(temp)
write.csv(temp,"allpca.csv")
```

```
#test for the assumption made above
#read data
dfall=read.csv('C:/Users/mac/Desktop/dataanalysis-R/class/final/csv/alldata.csv',header = TRUE, sep=',',na.strings = 'NA')
head(dfall)
```

```
##   city year treat urban       income        exp  labor
## 1 晋中 2013     1     1  14312.95528   4949.626 105000
## 2 运城 2013     1     1  2803.949366  -4550.336  78000
## 3 临汾 2013     1     1  8742.280394   1391.968 114000
## 4 金华 2013     1     1  11679.43847 138164.820 195000
## 5 衢州 2013     1     1 -6573.545511  73904.590 111000
## 6 萍乡 2013     1     1  8448.001378  99272.804 108000
```

```
#data cleaning and processing and format adjusting
dfall$post=ifelse(dfall$year>2014,1,0)
dfall=subset(dfall,income>0)
sapply(dfall,class)
```

```
##        city        year       treat       urban      income         exp
## "character"   "integer"   "integer"   "integer" "character"   "numeric"
##       labor        post
##   "integer"   "numeric"
```

```
dfall=as.data.frame(lapply(dfall[,2:7],as.numeric))
sapply(dfall,class)
```

```
##      year     treat     urban    income       exp     labor
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
```

```
#linear regression
dfall$post=ifelse(dfall$year>2014,1,0)
dfall$'labor_ind'=dfall$labor/10000
reg.urban_labor=lm(labor_ind~post*treat,data=subset(dfall,urban==1))
reg.urban_labor
```

```
##
## Call:
## lm(formula = labor_ind ~ post * treat, data = subset(dfall, urban ==
##     1))
##
## Coefficients:
## (Intercept)         post        treat   post:treat
##      13.807        1.711        3.314       -1.003
```

```
reg.rural_labor=lm(labor_ind~post*treat,data=subset(dfall,urban==0))
reg.rural_labor
```

```
##
## Call:
## lm(formula = labor_ind ~ post * treat, data = subset(dfall, urban ==
##     0))
##
## Coefficients:
## (Intercept)         post        treat   post:treat
##      10.180        6.487        8.813        2.053
```

```
#result shows that the construction of HSR had deteriorated the position of labor force in citi
es
#but did enhance the labor force of rural area
```

```
#linear regression
dfall$'income_ind'=dfall$income/1000
dfall$'exp_ind'=dfall$exp/1000
lrreg.urban=lm(income_ind~post+treat+labor_ind+exp_ind,data=subset(dfall,urban==1))
lrreg.urban
```

```
##
## Call:
## lm(formula = income_ind ~ post + treat + labor_ind + exp_ind,
##     data = subset(dfall, urban == 1))
##
## Coefficients:
## (Intercept)         post        treat    labor_ind      exp_ind
##    0.846653     5.133128     2.218272     0.568893    -0.002562
```

```
lrreg.rural=lm(income_ind~post+treat+labor_ind+exp_ind,data=subset(dfall,urban==0))
lrreg.rural
```

```
##
## Call:
## lm(formula = income_ind ~ post + treat + labor_ind + exp_ind,
##     data = subset(dfall, urban == 0))
##
## Coefficients:
## (Intercept)         post        treat    labor_ind      exp_ind
##     4.23677      8.20182     -0.26143     -0.23813      0.01746
```

```
#result shows that the relationship between labor force and income per capita is positive in ur
ban area and negative in rural area
```

```
stargazer(lrreg.urban, lrreg.rural, type="text",title = "Reason Analtsis",covariate.labels = c(
'POST','HSR','labor'), dep.var.labels = c("urban VS rural"),omit.stat = c("ser",'rsq',"adj.rsq"
))
```

```
##
## Reason Analtsis
## ===========================================================
##                          Dependent variable:
##              ---------------------------------------------
##                             urban VS rural
##                        (1)                     (2)
## -----------------------------------------------------------
## POST                5.133***                8.202***
##                      (1.911)                 (2.583)
##
## HSR                   2.218                  -0.261
##                      (2.026)                 (3.096)
##
## labor               0.569***                -0.238**
##                      (0.126)                 (0.095)
##
## exp_ind              -0.003                  0.017**
##                      (0.003)                 (0.008)
##
## Constant              0.847                   4.237
##                      (2.145)                 (2.736)
##
## -----------------------------------------------------------
## Observations           491                     406
## F Statistic  14.138*** (df = 4; 486) 4.162*** (df = 4; 401)
## ===========================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

```
#the table above represent that the construction of HSR did not lead to the decrease of rural labor force as we thought, but had a positive effect on that. However, as the rural labor force index is negatively correlated with its per capita income, it is understandable that the opening of HSR leads to the decrease of rural per capita income
#With regard to the conclusions above, we think that due to the accelerated urbanization process in rural areas, a large number of migrant workers have returned to the rural areas to work and live
#However, due to the imbalance of the rural labor force level, the increased labor force has brought about a decrease in per capita  and that is the reason why the increase in the rural labor force had decreased its personal income
```