STAT306 PROJECT

04.04.2017

Yumeng Chen	35365148
Rachel Gong	32282148
Xinwei Kuang	29223147
Shiyang Li	48753140
Yubin Lyu	47341145

Abstract

The purpose of this study is to form a prediction equation for the GDP per capita of 36 cities in China in 2014 based on the model generated from data of 2013. Explanatory variables being considered are listed in the following table. Region is a categorical variable with 7 categories of regions of China (i.e. North(N), South(S), East(E), Center(C), Northeast(NE), Southwest(SW), Northwest(NW)).

Variable	Description
pop	Total Population (year-end) (10^6 persons)
wage	Average Wage of Staff and Workers (10^6yuan)
stu	Number of Students Enrolment of Regular Institutions of Higher Education (10000 persons)
traffic	Passenger Traffic (10^7 persons)
sales	Total Retail Sales of Consumer Goods (10^4 million yuan)
hos	Number of Hospitals and Health Centers (unit)
region	The region of the city
asp	Average Selling Price of Commercialized Buildings (10^3yuan/sq.m)
sdh	Savings Deposit of Households, Balance at Year-end (10^5 million yuan)

Main Conclusion

Firstly, we use residual plots to see whether there is a pattern or heteroscedasticity in the plots, and use log term to eliminate these patterns. Then, we use exhausted algorithm to do variable selection to find which explanatory variables are useless. Finally, we use cross validation to compare the result between the model with smallest CP value and the model with largest adjusted R-squared value, then we could get the best fit model. The predicted equation we found is:

$$\underline{gc = 1.312622 + 0.341839*pop - 0.050425*stu + 0.016597*traffic + 0.546327*sales - 1.361474*sdh + 2.047223*iC + 2.273954*iN}$$

In the model above, dummy variables iC and iN indicate a binary variable that is 1 if the region variable has value of C and N respectively and 0 otherwise. With other explanatory variables held fixed, iC adds on average 2.047223 to gc, and iN adds on average 2.27395 to gc compared with South(baseline). With other explanatory variables held fixed, one more 106 persons of pop adds on average 0.341839 to gc, one more 104 persons of stu adds on average -0.05425 to gc, one more 107 persons of traffic adds on average 0.016597 to gc, one more 104 million yuan of sales adds on average 0.546327 to gc, one more 105 million yuan of sdh adds on average -1.361474 to gc.

Description of Data

In this study, the data was collected for GDP per capita of 36 cities in China in the year of 2013, and the data source is from the official websites of National Bureau of Statistics of China. The explanatory variables pop, wage, stu, traffic, sales, asp and sdh were in units of 10000 persons, yuan,10000 persons,10000 persons,100 million yuan, yuan/sq.m and 100 million yuan respectively, and we decided to change these units into 106 persons, 106 yuan,10000 persons,107 persons,104 million yuan, 103 yuan/sq.m and 105 million yuan because that made their residual plots look better.

The response variable gc represent the GDP per capita in units of $100s, and the sample size n=36.

Data Analysis and Results

<u>hos residual plot vs. logHos residual plot</u>

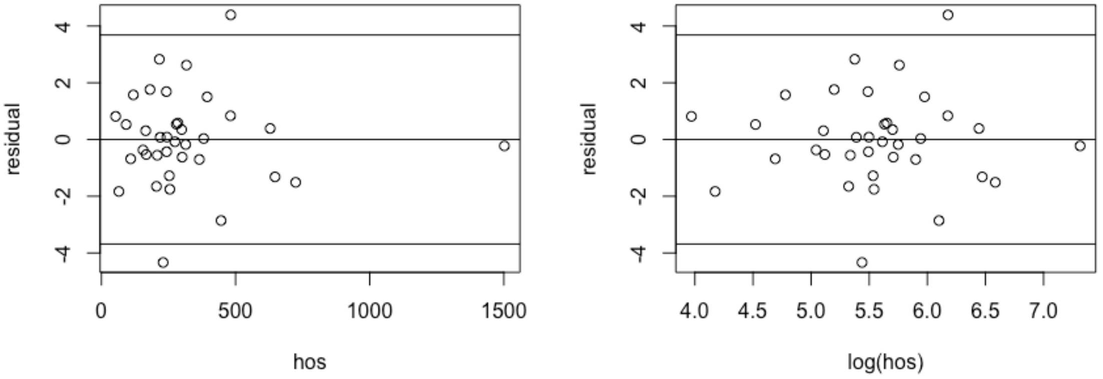
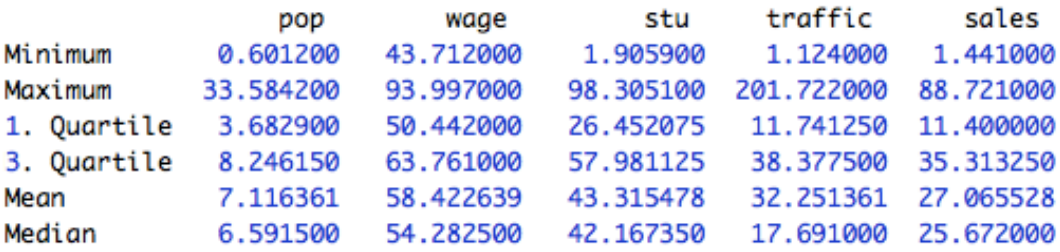


Figure 1

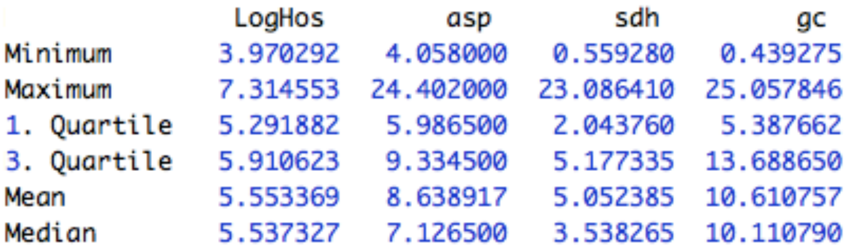
The residual plot of hos (left) and the residual plot of logHos (right)

Figure 1 suggests that explanatory variables hos need to be transformed to log(hos) due to the presence of outlier.

<u>Univariate summary statistics:</u>



```
                   pop        wage         stu      traffic       sales
Minimum       0.601200   43.712000    1.905900     1.124000    1.441000
Maximum      33.584200   93.997000   98.305100   201.722000   88.721000
1. Quartile   3.682900   50.442000   26.452075    11.741250   11.400000
3. Quartile   8.246150   63.761000   57.981125    38.377500   35.313250
Mean          7.116361   58.422639   43.315478    32.251361   27.065528
Median        6.591500   54.282500   42.167350    17.691000   25.672000
```



```
                LogHos         asp         sdh          gc
Minimum       3.970292    4.058000    0.559280    0.439275
Maximum       7.314553   24.402000   23.086410   25.057846
1. Quartile   5.291882    5.986500    2.043760    5.387662
3. Quartile   5.910623    9.334500    5.177335   13.688650
Mean          5.553369    8.638917    5.052385   10.610757
Median        5.537327    7.126500    3.538265   10.110790
```

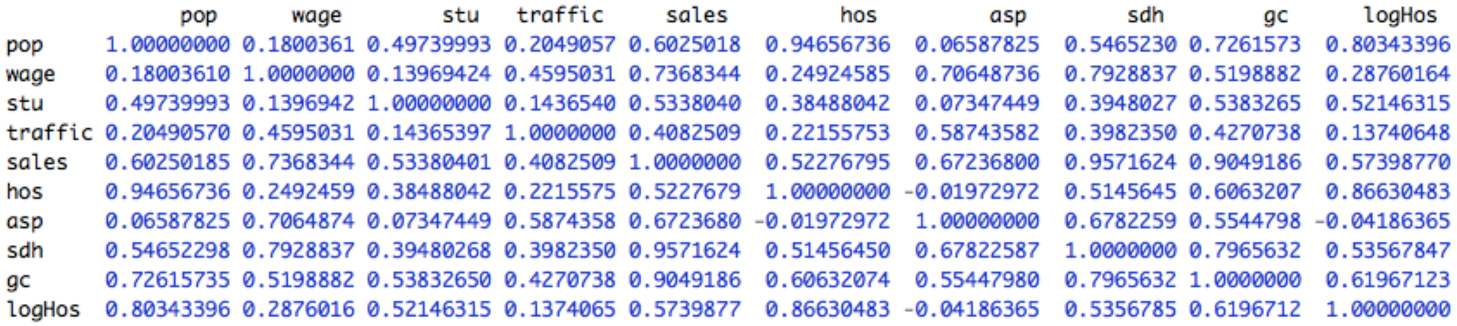
Table 1

Frequency table for Region:

C	E	N	NE	NW	S	SW
4	6	6	4	5	5	6

Table 2

Sample Correlations:



```
               pop      wage        stu   traffic     sales         hos         asp       sdh        gc      logHos
pop     1.00000000 0.1800361 0.49739993 0.2049057 0.6025018  0.94656736  0.06587825 0.5465230 0.7261573  0.80343396
wage    0.18003610 1.0000000 0.13969424 0.4595031 0.7368344  0.24924585  0.70648736 0.7928837 0.5198882  0.28760164
stu     0.49739993 0.1396942 1.00000000 0.1436540 0.5338040  0.38488042  0.07347449 0.3948027 0.5383265  0.52146315
traffic 0.20490570 0.4595031 0.14365397 1.0000000 0.4082509  0.22155753  0.58743582 0.3982350 0.4270738  0.13740648
sales   0.60250185 0.7368344 0.53380401 0.4082509 1.0000000  0.52276795  0.67236800 0.9571624 0.9049186  0.57398770
hos     0.94656736 0.2492459 0.38488042 0.2215575 0.5227679  1.00000000 -0.01972972 0.5145645 0.6063207  0.86630483
asp     0.06587825 0.7064874 0.07347449 0.5874358 0.6723680 -0.01972972  1.00000000 0.6782259 0.5544798 -0.04186365
sdh     0.54652298 0.7928837 0.39480268 0.3982350 0.9571624  0.51456450  0.67822587 1.0000000 0.7965632  0.53567847
gc      0.72615735 0.5198882 0.53832650 0.4270738 0.9049186  0.60632074  0.55447980 0.7965632 1.0000000  0.61967123
logHos  0.80343396 0.2876016 0.52146315 0.1374065 0.5739877  0.86630483 -0.04186365 0.5356785 0.6196712  1.00000000
```

Table 3

Since the correlation between logHos and pop is 0.8034 which shows that these two variables are highly correlated, we may need to select only one explanatory variable between these two; but we do not need to delete any variable here. We will use “exhaustive” selection to select useful variables and delete useless variables later.

Scatter plots:

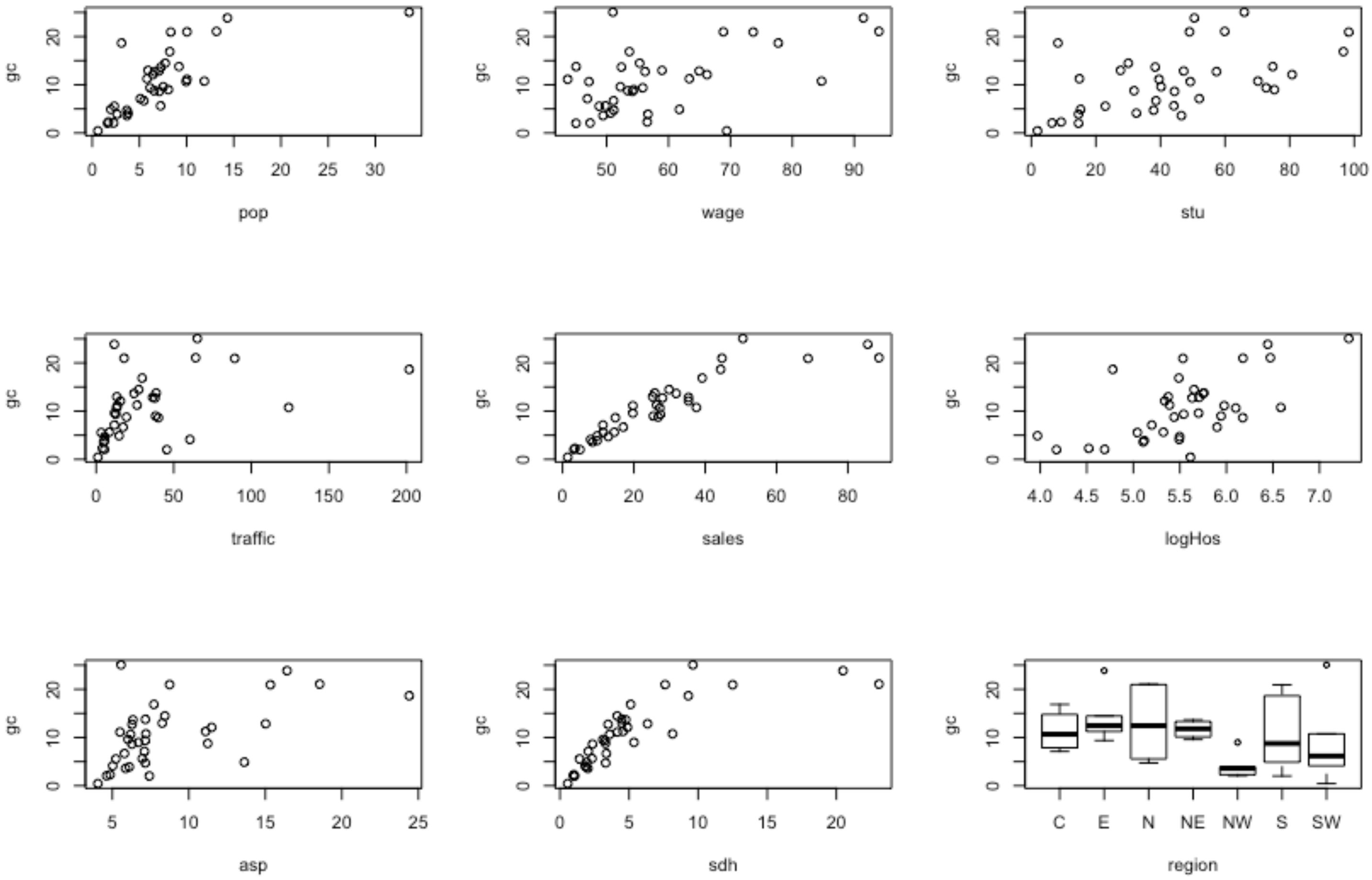


Figure 2
Scatter plots for gc vs every individual numerical explanatory variable (first eight graphs)
and boxplot of gc vs categorical variable region (last graph)

Residual plots:

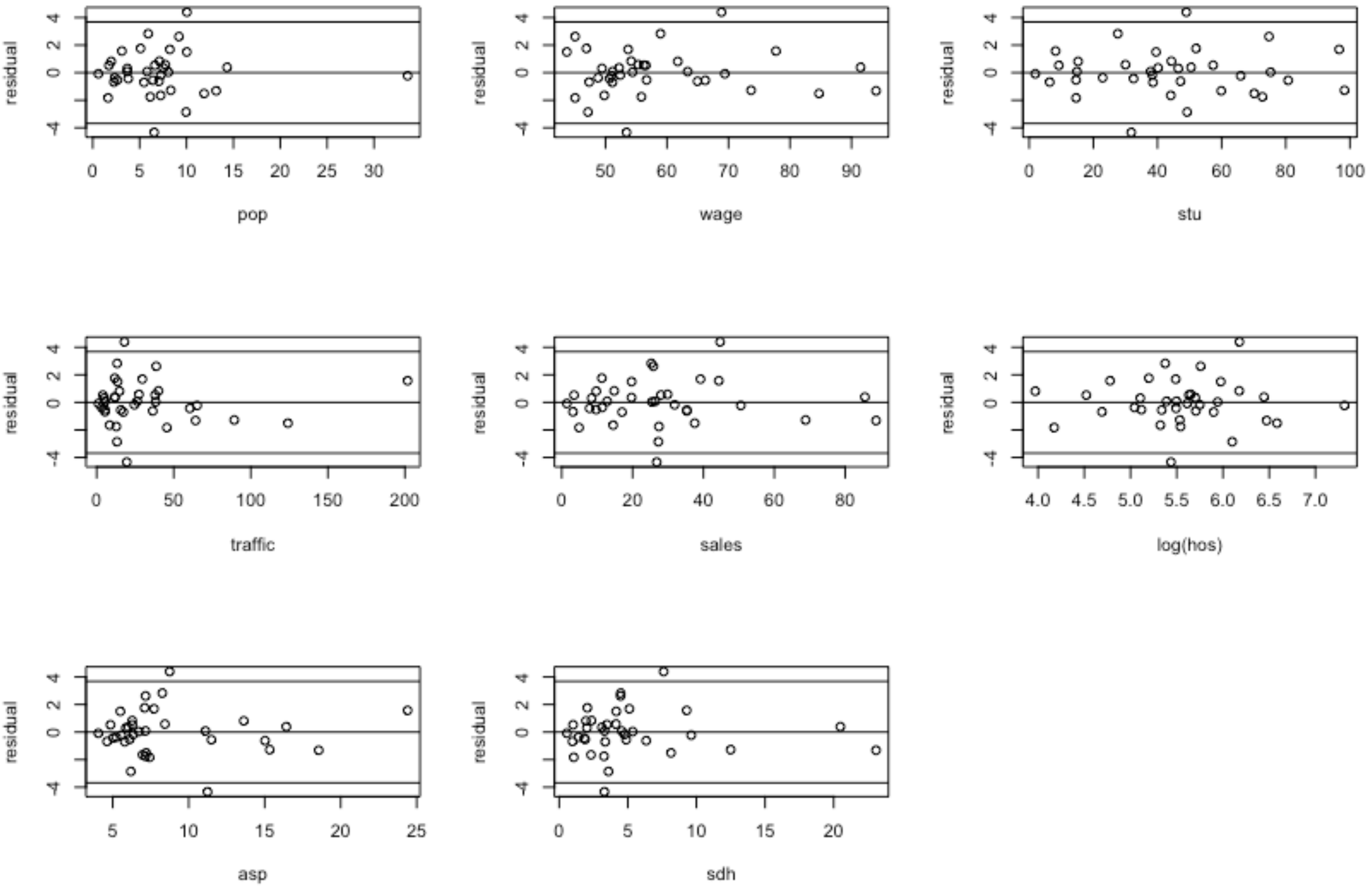


Figure 3
Residual plot for each explanatory variable
(we use logHos instead of hos due to the transformation of this variable)

The residual plots for wage, stu, and sales are homoscedastic and there is no specific pattern, so for those three explanatory variables, we didn't need to do any transformation. As for pop, even there is an outlier, other data points are in homoscedastic model, so we didn't transform pop. Residual plots for sales, asp, and sdh are a little bit left-skewed, so we tried to transform those three variables to their log, square root, and quadratic forms. However, even though the residual plots looked better with transformations, the adjusted R^2 got smaller when we were fitting the models. Hence, we finally decided to only transform hos to its log form, and kept all other explanatory variables remaining their original forms.

QQ-plot:

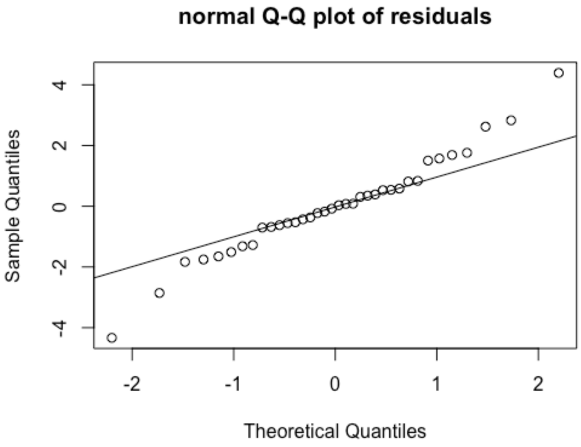
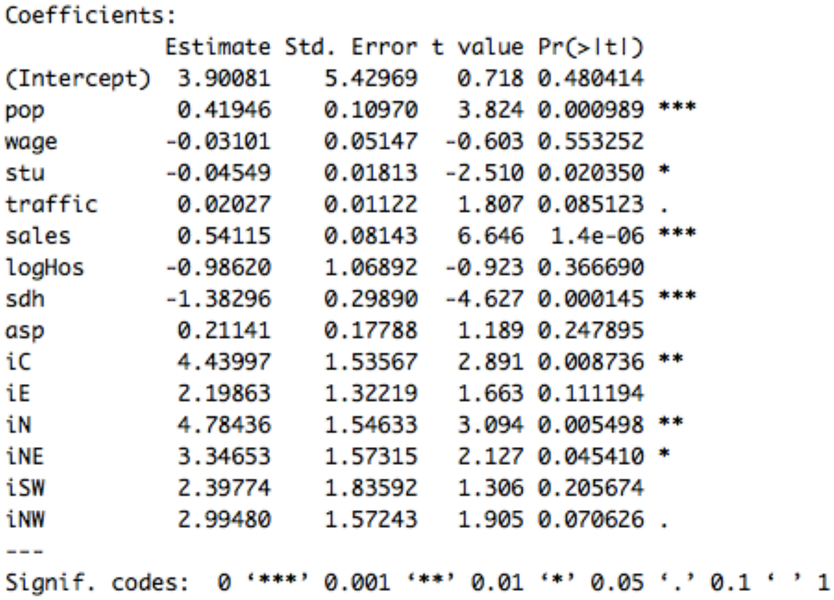


Figure 4
Normal Q-Q plot of residuals

The Q-Q plot looks quite good because most scatter points in the middle (from -1 to 1) are along the line, and there is no obvious upward or downward bending for the lower points and higher points.

First time fitting model:
Next, we fitted a multiple regression model with the explanatory variables as given (pop, wage, stu, traffic, sales, log(hos), sdh, asp, iC, iE, iN, iNE, iSW, iNW). After setting all dummy variables as baselines respectively, we found that overall significance is the highest (the same adj R^2) when we choose South as the baseline, so we set South as the baseline, and the following data is the summary of this fit.

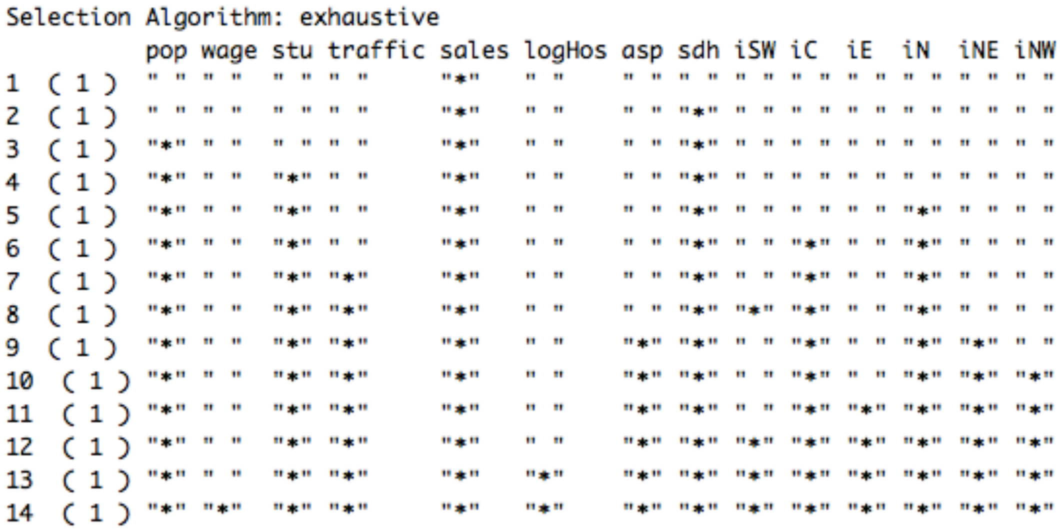


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.90081    5.42969   0.718 0.480414
pop          0.41946    0.10970   3.824 0.000989 ***
wage        -0.03101    0.05147  -0.603 0.553252
stu         -0.04549    0.01813  -2.510 0.020350 *
traffic      0.02027    0.01122   1.807 0.085123 .
sales        0.54115    0.08143   6.646  1.4e-06 ***
logHos      -0.98620    1.06892  -0.923 0.366690
sdh         -1.38296    0.29890  -4.627 0.000145 ***
asp          0.21141    0.17788   1.189 0.247895
iC           4.43997    1.53567   2.891 0.008736 **
iE           2.19863    1.32219   1.663 0.111194
iN           4.78436    1.54633   3.094 0.005498 **
iNE          3.34653    1.57315   2.127 0.045410 *
iSW          2.39774    1.83592   1.306 0.205674
iNW          2.99480    1.57243   1.905 0.070626 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 4
Data summary table with setting iS as the baseline

Variable selection

Then we used exhaustive selection method to do variable selection. The following table is the result of this kind of selection:



```
Selection Algorithm: exhaustive
          pop wage stu traffic sales logHos asp sdh iSW iC  iE  iN  iNE iNW
1  ( 1 )  " " " "  " " " "     "*"   " "    " " " " " " " " " " " " " " " "
2  ( 1 )  " " " "  " " " "     "*"   " "    " " "*" " " " " " " " " " " " "
3  ( 1 )  "*" " "  " " " "     "*"   " "    " " "*" " " " " " " " " " " " "
4  ( 1 )  "*" " "  "*" " "     "*"   " "    " " "*" " " " " " " " " " " " "
5  ( 1 )  "*" " "  "*" " "     "*"   " "    " " "*" " " " " " " "*" " " " "
6  ( 1 )  "*" " "  "*" " "     "*"   " "    " " "*" " " "*" " " "*" " " " "
7  ( 1 )  "*" " "  "*" "*"     "*"   " "    " " "*" " " "*" " " "*" " " " "
8  ( 1 )  "*" " "  "*" "*"     "*"   " "    " " "*" "*" "*" " " "*" " " " "
9  ( 1 )  "*" " "  "*" "*"     "*"   " "    "*" "*" " " "*" " " "*" "*" " "
10  ( 1 ) "*" " "  "*" "*"     "*"   " "    "*" "*" " " "*" " " "*" "*" "*"
11  ( 1 ) "*" " "  "*" "*"     "*"   " "    "*" "*" " " "*" "*" "*" "*" "*"
12  ( 1 ) "*" " "  "*" "*"     "*"   " "    "*" "*" "*" "*" "*" "*" "*" "*"
13  ( 1 ) "*" " "  "*" "*"     "*"   "*"    "*" "*" "*" "*" "*" "*" "*" "*"
14  ( 1 ) "*" "*"  "*" "*"     "*"   "*"    "*" "*" "*" "*" "*" "*" "*" "*"
```

Table 5
Table of exhaustive selection result

We found that Line 7 has the smallest CP value (8.251486); and Line 11 has the largest adjusted R-squared (0.9427609). So, we defined two models for cross validation according to these two lines. Model 1 is the model with the smallest CP value, and the explanatory variables are: pop, stu, traffic, sales, sdh, iC, and iN; Model 2 is the model with the largest adjusted R-squared, and the explanatory variables are: pop, stu, traffic, sales, asp, sdh, iE, iC, iN, iNE, and iNW.

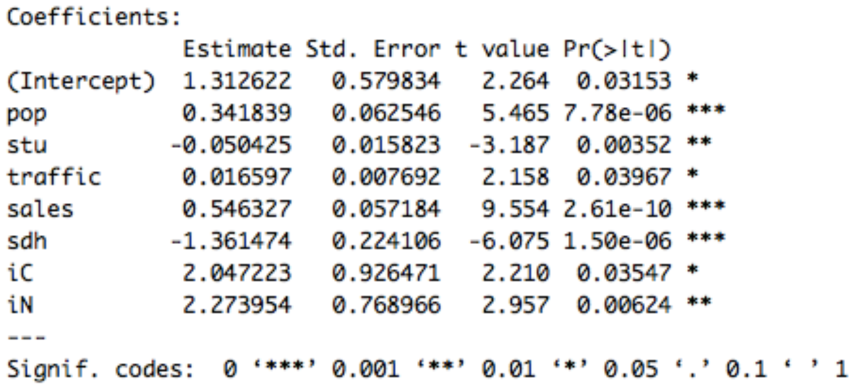
Cross-validation and out-of-sample comparisons:

statistic\model	1	2
adjusted R2	0.9399	0.9428
residual SD	1.593	1.555
RMSE(leave-one-out)	2.1577	2.1712
RMSE(5-fold)	2.2056	2.3155

Table 6
Summary cross-validation table of selected model 1 and model 2

This table shows the comparisons of the two selected models. Although model 2 has greater adjusted R2 and smaller residual SD, smaller RMSE values of leave-one-out and 5-fold suggest that model 1 is a better model than model 2.

The summary of the model we get is:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.312622   0.579834   2.264  0.03153 *  
pop          0.341839   0.062546   5.465 7.78e-06 ***
stu         -0.050425   0.015823  -3.187  0.00352 ** 
traffic      0.016597   0.007692   2.158  0.03967 *  
sales        0.546327   0.057184   9.554 2.61e-10 ***
sdh         -1.361474   0.224106  -6.075 1.50e-06 ***
iC           2.047223   0.926471   2.210  0.03547 *  
iN           2.273954   0.768966   2.957  0.00624 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 7
Summary table for Model 1
(with explanatory variables pop, stu, traffic, sales, sdh, iC, and iN)

From the information above, the final model generated is:
gc = 1.312622 + 0.341839*pop - 0.050425*stu + 0.016597*traffic + 0.546327*sales - 1.361474*sdh + 2.047223*iC + 2.273954*iN

Brief Discussion

In conclusion, we have found a best-fitting model and residual plot that can form a prediction equation for the GDP per capita of 36 cities in China in 2014.
The resulting beast prediction equation is:

*gc = 1.312622 + 0.341839*pop - 0.050425*stu + 0.016597*traffic + 0.546327*sales - 1.361474*sdh + 2.047223*iC + 2.273954*iN*

Adding quadratic term didn’t show any improvement, because after we added the quadratic term to pop, wage, sdh and asp, the Adjusted R-squared decreased. The Adjusted R-squared decreased from 0.9404 to 0.9342. Therefore, the quadratic model is worse than the original model. Meanwhile, adding log of hos did make improvement. The adjusted R-squared increased from 0.9196 to 0.9197.
The prediction equation can be tested to see how well they predict GDP per capital in unit of $100 after using the real data from 2014. However, for predicting GDP per capital in future years, the predictions may not be precise because there may exist other explanatory variables correlated to the ones here, and regression coefficients in our model will also change.

Contribution:

1. All the members in our team are friends.
2. Name of authors (ordering by alphabetical by surname):

Yumeng Chen	35365148
Rachel Gong	32282148
Xinwei Kuang	29223147
Shiyang Li	48753140
Yubin Lyu	47341145

3. Contribution on this project:

Shiyang and Rachel raised the main idea of this project, and we all discussed together for details.
Yubin and Xinwei wrote the main R code for this project.
Yumeng and Shiyang found some problems in R code and revised some of them.
For the project part, we all contribute and work hard in this final project:
Yumeng organized the plots in this project;
Yumeng and Shiyang wrote the Data Analysis and Results part of this project;
Yubin and Xinwei wrote abstract part and cross-validation part of this project;
Rachel, Yubin and Xinwei wrote the discussion part of this project.