STAT306 PROJECT

04.04.2017

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

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 are pop, wage, stu, traffic, sales, hos, region, asp, sdh. 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)).

Main Conclusion

Firstly, we use residual plots to see whether there is a pattern or heteroscedasticity in the plots, and use quadratic term or 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:

*gc = 22.410486 - 2.163781*asp + 0.551688*sales - 0.307262*pop + 0.112715*wage - 0.048447*stu - 3.159237*logHos + 0.142104*I(asp^2) - 0.029534*I(sdh^2) + 2.180527*iC + 2.089837*iN - 0.020519*asp:sales*

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 2.180527 to gc, and iN adds on average 2.089837 to gc compared with Northwest(baseline). With other explanatory variables held fixed, one more 1000 yuan/sq.m asp adds on average - 2.163781 to gc, one more 10000 million yuan of sales adds on average 0.551688 to gc, one more 1000000 persons of pop adds on average -0.307262 to gc, one more 1000000 yuan adds on average 0.112715 to gc, one more 10000 persons of stu adds on average - 0.048447 to gc, one more unit of logHos adds on average - 3.159237 to gc, one more unit of I(asp^2) adds on average 0.142104 to gc, one more unit of I(sdh^2) adds on average - 0.029534 to gc.

Description of Data

Variable	Description
pop	Total Population (year-end) (10^4 persons)
wage	Average Wage of Staff and Workers (yuan)

stu	Number of Students Enrolment of Regular Institutions of Higher Education (10^4 persons)
traffic	Passenger Traffic (10^4 persons)
sales	Total Retail Sales of Consumer Goods (100 million yuan)
hos	Number of Hospitals and Health Centers (unit)
region	The region of the city
asp	Average Selling Price of Commercialized Buildings (yuan/sq.m)
sdh	Savings Deposit of Households, Balance at Year-end (100 million yuan)

Table 1
Summary of data short names and units

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 originally in units of 10000 persons, yuan,10000 persons,10000 persons,100 million yuan, yuan/sq.m and 100 million yuan respectively, and we transformed these units into 1000000 persons, 1000000 yuan,10000 persons,10000000 persons,10000 million yuan, 1000 yuan/sq.m and 100000 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

Scatter plot

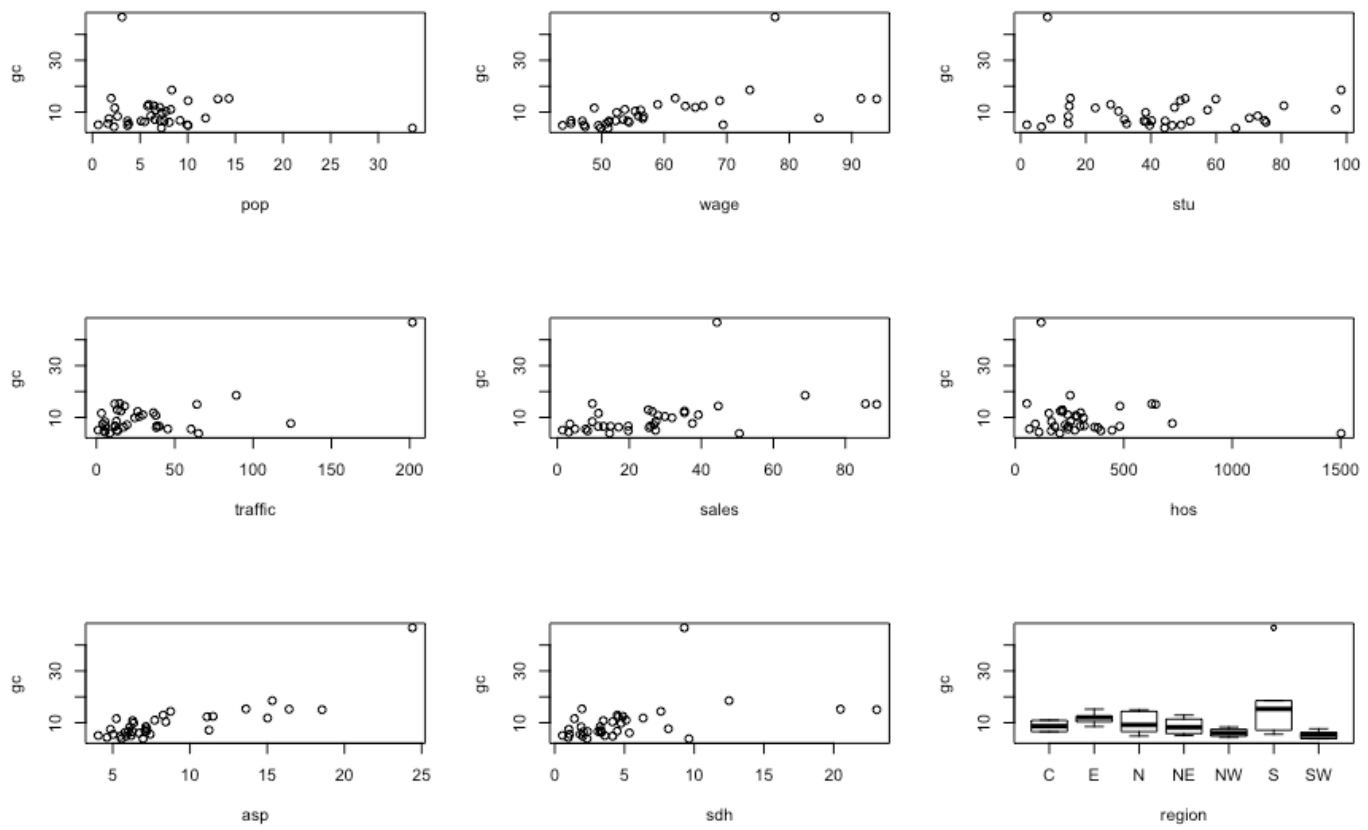

Figure 1

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

Residual plot

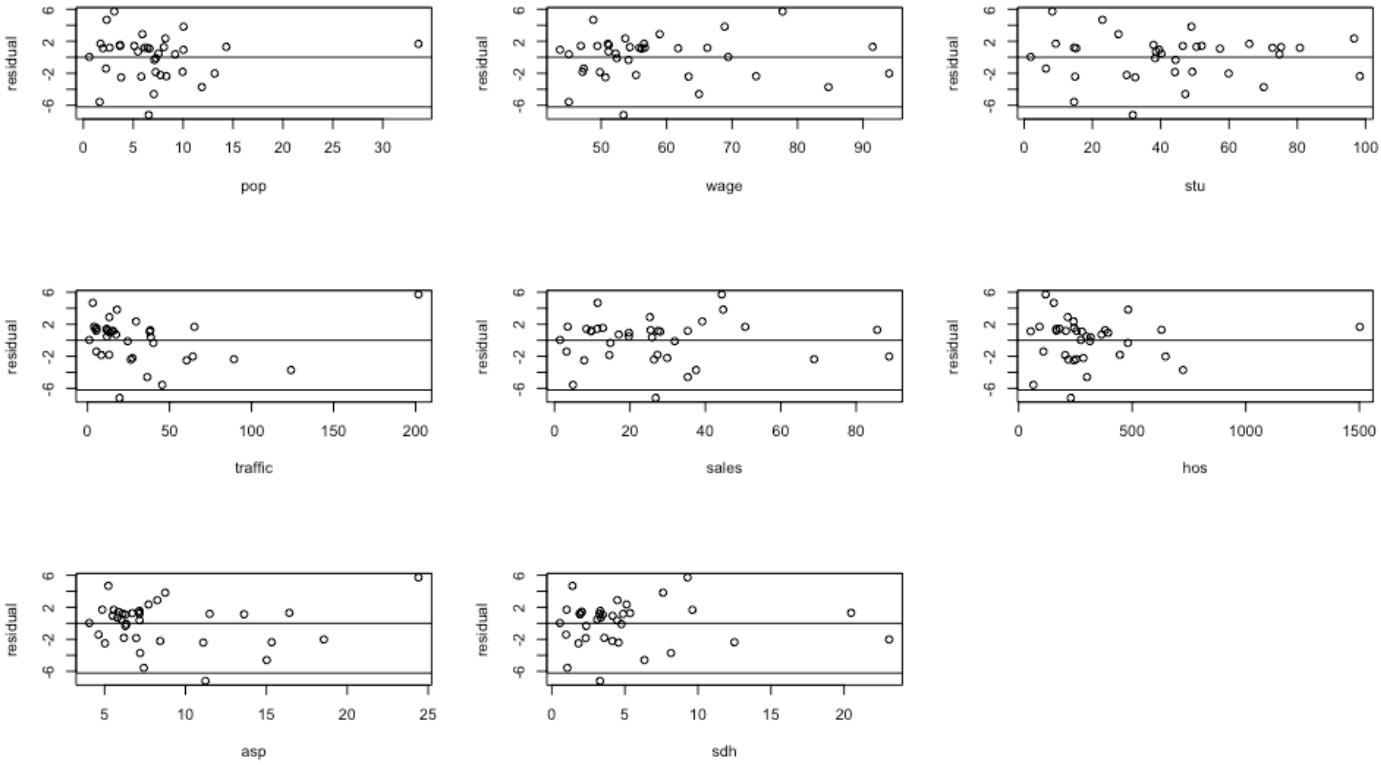

Figure 2
Residual plot for each explanatory variable

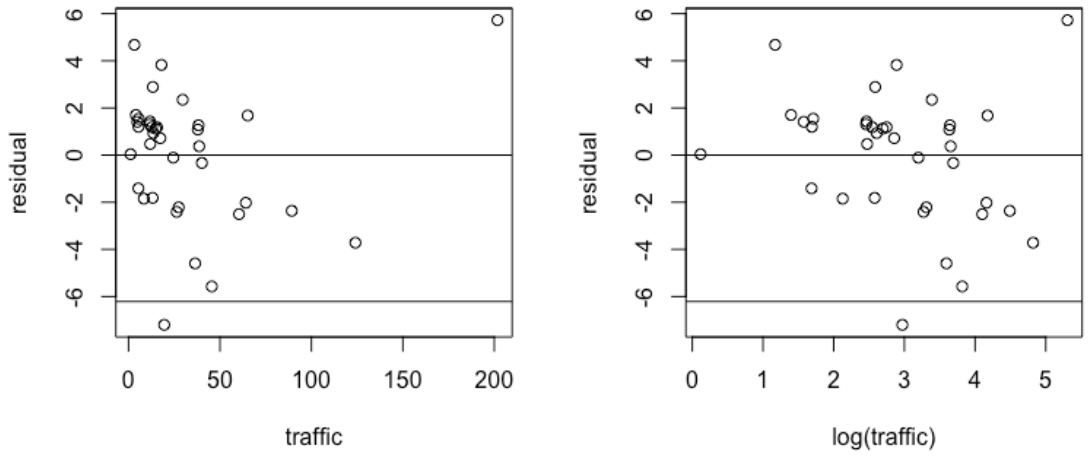

Figure 3
The residual plot of traffic (left) and the residual plot of logTraffic (right)

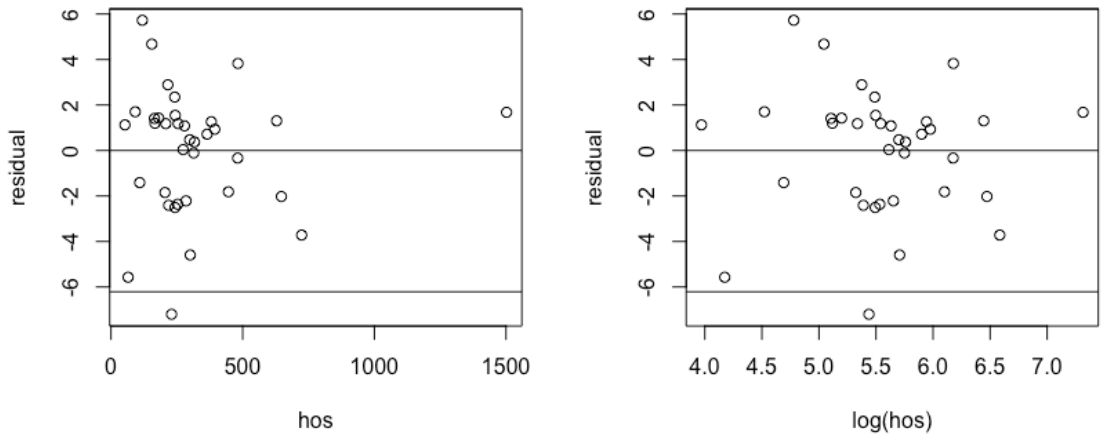

Figure 4
The residual plot of hos (left) and the residual plot of logHos (right)

The residual plots for wage and stu are homoscedastic and there are no specific patterns, so for those two explanatory variables, we didn't do any transformations. However, for all other variables (pop, traffic, sales, hos, asp and sdh), there are some patterns. 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 forms. However, even though the residual plots looked better with transformations, the adjusted R- squared got smaller when we were fitting the models. Figure 3 and Figure 4 suggest that explanatory variables hos and traffic need to be transformed to log(hos) and log(traffic) due to the presence of outliers and heteroscedastic patterns because when we add log to hos and traffic, the adjusted R-squared increase and patterns are removed. Then, we tried to add quadratic terms to sales, asp, and sdh and the adjusted R-squared increased to 0.9442, so we kept the transformations.
In short, we transformed traffic and hos to their log forms, and sales, asp and sdh to their quadratic forms, while keeping wage and stu remaining their original form.

Univariate summary statistics:

```
      pop              wage             stu             traffic      
 Min.   : 0.6012   Min.   :43.71   Min.   : 1.906   Min.   :  1.124  
 1st Qu.: 3.6829   1st Qu.:50.44   1st Qu.:26.452   1st Qu.: 11.741  
 Median : 6.5915   Median :54.28   Median :42.167   Median : 17.691  
 Mean   : 7.1164   Mean   :58.42   Mean   :43.315   Mean   : 32.251  
 3rd Qu.: 8.2462   3rd Qu.:63.76   3rd Qu.:57.981   3rd Qu.: 38.377  
 Max.   :33.5842   Max.   :94.00   Max.   :98.305   Max.   :201.722  
     sales             hos              asp              sdh        
 Min.   : 1.441   Min.   :  53.0   Min.   : 4.058   Min.   : 0.5593  
 1st Qu.:11.400   1st Qu.: 199.0   1st Qu.: 5.987   1st Qu.: 2.0438  
 Median :25.672   Median : 254.0   Median : 7.127   Median : 3.5383  
 Mean   :27.066   Mean   : 320.6   Mean   : 8.639   Mean   : 5.0524  
 3rd Qu.:35.313   3rd Qu.: 369.0   3rd Qu.: 9.335   3rd Qu.: 5.1773  
 Max.   :88.721   Max.   :1502.0   Max.   :24.402   Max.   :23.0864  
       gc              logHos         logTraffic        salesSQ      
 Min.   : 3.806   Min.   :3.970   Min.   :0.1169   Min.   :   2.076  
 1st Qu.: 5.925   1st Qu.:5.292   1st Qu.:2.4631   1st Qu.: 129.962  
 Median : 7.567   Median :5.537   Median :2.8729   Median : 659.088  
 Mean   : 9.879   Mean   :5.553   Mean   :2.9340   Mean   :1159.375  
 3rd Qu.:11.928   3rd Qu.:5.911   3rd Qu.:3.6475   3rd Qu.:1247.026  
 Max.   :46.704   Max.   :7.315   Max.   :5.3069   Max.   :7871.416  
     aspSQ             sdhSQ          aspAndsdhAndsalesSQ
 Min.   : 16.47   Min.   :  0.3128   Min.   :   36.7    
 1st Qu.: 35.84   1st Qu.:  4.1771   1st Qu.:  511.4    
 Median : 50.79   Median : 12.5224   Median : 1408.3    
 Mean   : 94.22   Mean   : 48.9711   Mean   : 2474.5    
 3rd Qu.: 88.17   3rd Qu.: 26.8156   3rd Qu.: 2680.3    
 Max.   :595.46   Max.   :532.9823   Max.   :16993.8    
```

Table 2

(Note: aspAndsdhAndsalesSQ is the square of the sum of asp, sdh, and sales)

Frequency table for Region:

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

Table 3

Sample Correlations:

	pop	wage	stu	traffic	sales	hos	asp	sdh
pop	1.00000000	0.1800361	0.49739993	0.2049057	0.6025018	0.94656736	0.06587825	0.5465230
wage	0.18003610	1.0000000	0.13969424	0.4595031	0.7368344	0.24924585	0.70648736	0.7928837
stu	0.49739993	0.1396942	1.00000000	0.1436540	0.5338040	0.38488042	0.07347449	0.3948027
traffic	0.20490570	0.4595031	0.14365397	1.0000000	0.4082509	0.22155753	0.58743582	0.3982350
sales	0.60250185	0.7368344	0.53380401	0.4082509	1.0000000	0.52276795	0.67236800	0.9571624
hos	0.94656736	0.2492459	0.38488042	0.2215575	0.5227679	1.00000000	-0.01972972	0.5145645
asp	0.06587825	0.7064874	0.07347449	0.5874358	0.6723680	-0.01972972	1.00000000	0.6782259
sdh	0.54652298	0.7928837	0.39480268	0.3982350	0.9571624	0.51456450	0.67822587	1.0000000
gc	-0.10291417	0.5511552	-0.08370726	0.6941914	0.4434592	-0.15762363	0.83552945	0.4141688
logHos	0.80343396	0.2876016	0.52146315	0.1374065	0.5739877	0.86630483	-0.04186365	0.5356785
logTraffic	0.41086899	0.3083658	0.42383800	0.8062082	0.5246926	0.34621546	0.49875978	0.4540052
salesSQ	0.49535901	0.7745707	0.38887419	0.3087603	0.9453815	0.45395205	0.64797077	0.9802634
aspSQ	0.03413759	0.6779098	-0.01609950	0.6694863	0.6101753	-0.02439381	0.97490050	0.6370764
sdhSQ	0.41303577	0.7619607	0.24964553	0.2469734	0.8517978	0.41146678	0.60643500	0.9531479
aspAndsdhAndsalesSQ	0.43801886	0.8011122	0.33386626	0.3596882	0.9343768	0.39965079	0.71707298	0.9819148

	gc	logHos	logTraffic	salesSQ	aspSQ	sdhSQ	aspAndsdhAndsalesSQ
pop	-0.10291417	0.80343396	0.4108690	0.4953590	0.03413759	0.4130358	0.4380189
wage	0.55115517	0.28760164	0.3083658	0.7745707	0.67790983	0.7619607	0.8011122
stu	-0.08370726	0.52146315	0.4238380	0.3888742	-0.01609950	0.2496455	0.3338663
traffic	0.69419139	0.13740648	0.8062082	0.3087603	0.66948628	0.2469734	0.3596882
sales	0.44345923	0.57398770	0.5246926	0.9453815	0.61017532	0.8517978	0.9343768
hos	-0.15762363	0.86630483	0.3462155	0.4539521	-0.02439381	0.4114668	0.3996508
asp	0.83552945	-0.04186365	0.4987598	0.6479708	0.97490050	0.6064350	0.7170730
sdh	0.41416877	0.53567847	0.4540052	0.9802634	0.63707639	0.9531479	0.9819148
gc	1.00000000	-0.18435974	0.4355088	0.3730554	0.89302706	0.3076243	0.4482104
logHos	-0.18435974	1.00000000	0.3066022	0.4717228	-0.06267166	0.4224225	0.4193494
logTraffic	0.43550878	0.30660224	1.0000000	0.3688887	0.48334533	0.2797890	0.3861774
salesSQ	0.37305543	0.47172275	0.3688887	1.0000000	0.59918816	0.9635158	0.9930908
aspSQ	0.89302706	-0.06267166	0.4833453	0.5991882	1.00000000	0.5743271	0.6781919
sdhSQ	0.30762431	0.42242246	0.2797890	0.9635158	0.57432711	1.0000000	0.9698703
aspAndsdhAndsalesSQ	0.44821040	0.41934937	0.3861774	0.9930908	0.67819190	0.9698703	1.0000000

Table 4

Since some variables are highly correlated according to Table 3 (for example, sales and sdh have a correlation value of 0.957), we may need to select only one variable among the correlated variables; but we do not need to delete any variable here. We will use "exhaustive" selection to select useful variables and delete useless variables later.

QQ-plot:

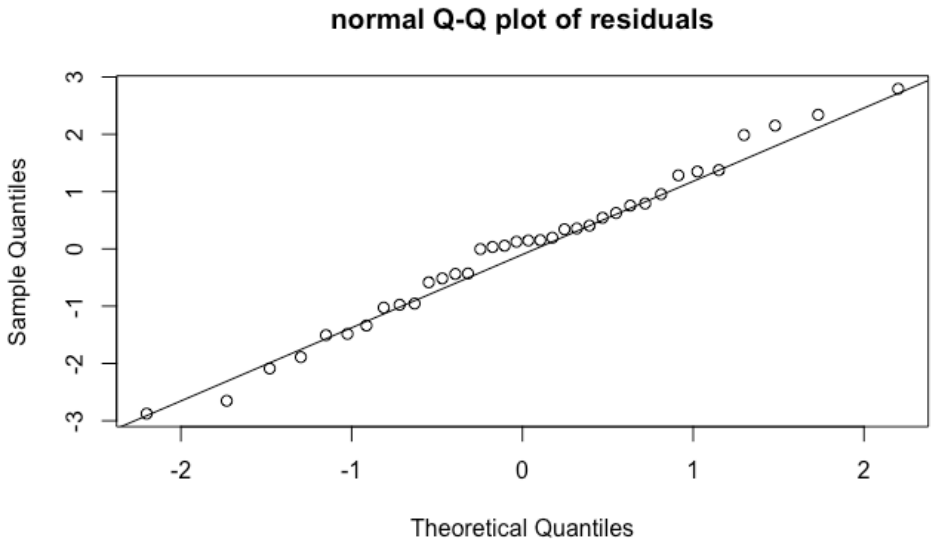

Figure 4
Normal Q-Q plot of residuals

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 (asp + sdh + sales)^2, pop, wage, stu, logTtraffic, I(sales^2), logHos, I(asp^2), I(sdh^2), iC, iE, iS, iNE, iSW, iN. After setting all dummy variables as baselines respectively, we found that overall significance is the highest (the same adj R^2) when we choose Northwest as the baseline, so we set Northwest as the baseline, and the following data is the summary of this fit.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.50105    8.41758   2.673   0.0174 *
asp         -1.83536    0.70758  -2.594   0.0203 *
sdh         -1.90667    2.02500  -0.942   0.3613
sales        0.84489    0.34095   2.478   0.0256 *
pop         -0.29916    0.22789  -1.313   0.2090
wage         0.12522    0.06770   1.850   0.0842 .
stu         -0.03568    0.02254  -1.583   0.1343
logTraffic   0.21992    0.58733   0.374   0.7133
I(sales^2)  -0.01788    0.01264  -1.415   0.1776
logHos      -3.57667    1.51365  -2.363   0.0321 *
I(asp^2)     0.13148    0.04465   2.945   0.0100 *
I(sdh^2)    -0.45063    0.25011  -1.802   0.0917 .
iC           1.87854    1.70775   1.100   0.2887
iE          -0.25919    1.76401  -0.147   0.8851
iS          -1.71036    2.51805  -0.679   0.5073
iNE          0.19515    1.58815   0.123   0.9038
iSW         -0.09767    1.33149  -0.073   0.9425
iN           2.32057    1.29749   1.789   0.0939 .
asp:sdh      0.11691    0.28653   0.408   0.6890
asp:sales   -0.04278    0.04482  -0.955   0.3549
sdh:sales    0.17849    0.10538   1.694   0.1110
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.65 on 15 degrees of freedom
Multiple R-squared:  0.9786,    Adjusted R-squared:  0.9501
F-statistic: 34.34 on 20 and 15 DF,  p-value: 4.221e-09
```

Table 5

Data summary table with setting iNW 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
         asp sdh sales pop wage stu logTraffic I(sales^2) logHos I(asp^2) I(sdh^2) iC  iE  iS  iNE iSW iN  asp:sdh asp:sales sdh:sales
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 ) "*" " " "*"  "*" "*" "*" " "          "*"       "*"     "*"      "*"      "*" " " " " "*" " " " " "*" " "      "*"       "*"
15 ( 1 ) "*" "*" "*"  "*" "*" "*" " "          "*"       "*"     "*"      "*"      "*" " " " " "*" " " " " "*" " "      "*"       "*"
16 ( 1 ) "*" "*" "*"  "*" "*" "*" " "          "*"       "*"     "*"      "*"      "*" " " " " "*" " " " " "*" "*"      "*"       "*"
17 ( 1 ) "*" "*" "*"  "*" "*" "*" "*"          "*"       "*"     "*"      "*"      "*" " " " " "*" " " " " "*" "*"      "*"       "*"
18 ( 1 ) "*" "*" "*"  "*" "*" "*" "*"          "*"       "*"     "*"      "*"      "*" " " " " "*" "*" " " "*" "*"      "*"       "*"
19 ( 1 ) "*" "*" "*"  "*" "*" "*" "*"          "*"       "*"     "*"      "*"      "*" "*" "*" "*" " " " " "*" "*"      "*"       "*"
20 ( 1 ) "*" "*" "*"  "*" "*" "*" "*"          "*"       "*"     "*"      "*"      "*" "*" "*" "*" "*" "*" "*" "*"      "*"       "*"
```

Table 5
Table of exhaustive selection result

We found that Line 11 has the smallest CP value (8.602614); and Line 15 has the largest adjusted R-squared (0.9609744). 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: asp, sales, pop, wage, stu, logHos, I(asp^2), I(sdh^2), iC, iN, and asp:sales; Model 2 is the model with the largest adjusted R-squared, and the explanatory variables are: asp, sdh, sales, pop, wage, stu, I(asp^2), I(sdh^2), logHos, I(sales^2), iC, iS, iN, asp:sales, and sdh:sales.

Cross-validation and out-of-sample comparisons:

statistic\model	1	2
adjusted R2	0.9572	0.961
residual SD	1.528	1.459
RMSE(leave-one-out)	3.090299	3.765148
RMSE(5-fold)	3.065682	4.280561

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) 22.410486   4.765739   4.702 8.85e-05 ***
asp         -2.163781   0.339811  -6.368 1.39e-06 ***
sales        0.551688   0.097652   5.650 8.11e-06 ***
pop         -0.307262   0.134730  -2.281  0.03175 *
wage         0.112715   0.044706   2.521  0.01874 *
stu         -0.048447   0.017566  -2.758  0.01094 *
logHos      -3.159237   0.884181  -3.573  0.00154 **
I(asp^2)     0.142104   0.016330   8.702 6.90e-09 ***
I(sdh^2)    -0.029534   0.012567  -2.350  0.02733 *
iC           2.180527   0.922253   2.364  0.02649 *
iN           2.089837   0.803744   2.600  0.01570 *
asp:sales   -0.020519   0.008741  -2.347  0.02749 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.528 on 24 degrees of freedom
Multiple R-squared:  0.9706,    Adjusted R-squared:  0.9572
F-statistic: 72.14 on 11 and 24 DF,  p-value: 1.237e-15
```

Table 6
Summary table for Model 1

From the information above, the final model generated is:

*gc = 22.410486 - 2.163781*asp + 0.551688*sales - 0.307262*pop + 0.112715*wage - 0.048447*stu - 3.159237*logHos + 0.142104*I(asp^2) - 0.029534*I(sdh^2) + 2.180527*iC + 2.089837*iN - 0.020519*asp:sales*

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 = 22.410486 - 2.163781*asp + 0.551688*sales - 0.307262*pop + 0.112715*wage - 0.048447*stu - 3.159237*logHos + 0.142104*I(asp^2) - 0.029534*I(sdh^2) + 2.180527*iC + 2.089837*iN - 0.020519*asp:sales*

Adding quadratic term did show improvement, because after we added the quadratic term to asp, sdh and sales, the Adjusted R-squared increased. The Adjusted R-squared increased from 0.8233 to 0.9373. Therefore, the quadratic model is better than the original model. Meanwhile, adding log of hos and traffic did make improvement. The adjusted R-squared increased from 0.9373 to 0.9442.

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.