Ming Xian
Professor Jeffrey Simonoff
STAT-UB 17
28 April, 2021

Regression Report of Two-way ANOVA

· Problem at hand & Why it's interesting to me

With the continuous progress of Chinese society, although the GDP and economic level of all provinces have significantly improved, regional development inequality among provinces has also been aggravated. In this report, I'm interested in finding out some possible factors which may explain the differences among the gross regional product of different provinces in China. Specifically, the two predicting factors I used are both categorical variables -- one is the geographical region of the province (three levels: east, central and west), and the other is the industry sector which is of the most labor in this province (three levels: manufacturing and construction sector, basic service and social service sector, and finance sector). Note that all my analysis on their relationship is based on the same time period of these variables. My decision of choosing these two categorical predictors has the following underlying logic. The geographical position of a region determines the degree of traffic convenience, thereby affecting economic development. The industry with the largest labor force in certain provinces indicates the dominant industry in that province, which would be interesting to dig into the relationship between industry and economic development.

· Description of the Dataset / Statistical Methodology

According to Wikipedia, gross regional product is "a monetary measure of the market value of all final goods and services produced in a region or subdivision of a country in a period (quarterly or yearly) of time". GRP per capita, as the name suggests, is calculated through dividing GRP by its population, measuring the economic output per person of each region. So, I've collected GRP per capita for 31 provinces in China (sample size: 31) in the year 2019, including 4 municipalities directly under the central government (Beijing, Tianjin, Shanghai, Chongqing), from the National Bureau of Statistics of China (https://data.stats.gov.cn/index.htm). For the division of regions, I divided all provinces into 3 parts (East, Central and West) by referring to the official standard, which can be found on this website: https://www.unicef.cn/en/figure-11-geographic-regions-china. For the other categorical predicting variable "industry with the largest number of employees" for each province, I found the data for 2019 on China Population and Employment

Statistics Yearbook Website (https://www.yearbookchina.com/navibooklist-n3020013208-1.html). For simplicity, I denoted this variable as "leading sector" in the following analysis. According to their classification, the urban employment population is mainly distributed in manufacturing, construction, basic services, social services and finance sectors. To simplify my analysis and regarding the similarities between some industries, I used three levels to classify the industry of most labor for each province: 1) manufacturing and construction (denoted as "m&c" in my dataset and following discussions for simplicity), 2) basic services and social services (denoted as "service" in my dataset and following discussions), 3) finance.

I want to examine the linear relationship between these two categorical factors (region & leading sector) and GRP per capita, with these two categorical factors treated as predicting variables and housing prices as the response variable, by using a two-way ANOVA model:

$$\mathbf{y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\varepsilon_{ijk}(random\ error),\ i=1,2,3,\ j=1,2,3,\ k=1,...,n_{ij}.}$$

Here α_i represents the main effect of the rows (region), β_j represents the main effect of the columns (leading sector), and $(\alpha\beta)_{ij}$ represents an interaction effect of these rows and columns. Whether we need this interaction term will be examined later. The appropriateness of this relationship can be tested by using Minitab 19.

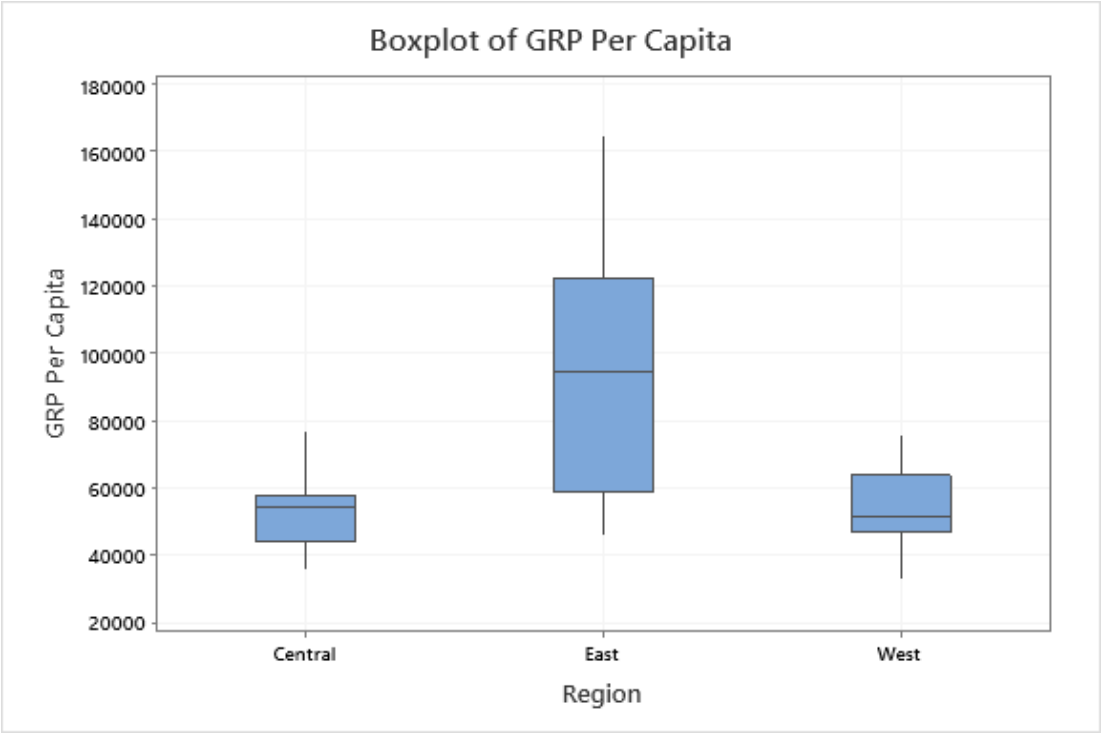
· Discussions of the Results

First let's take a look at the descriptive statistics.

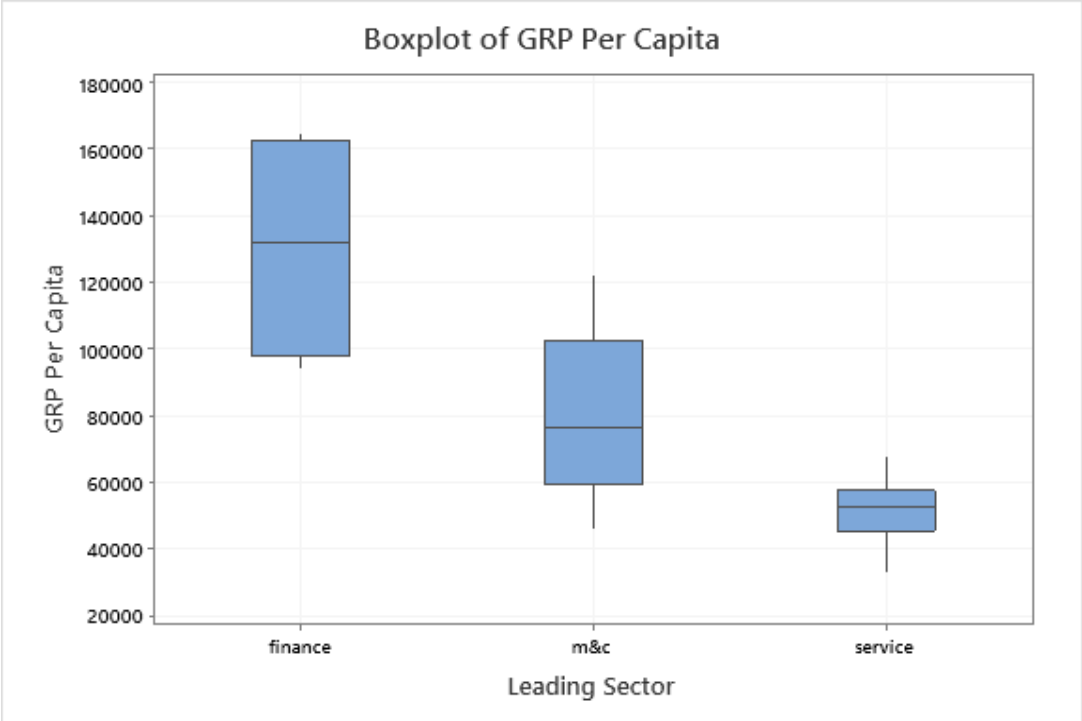
Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
GRP Per Capita	31	0	69139	5858	32615	32995	47944	57067	76712	164563

A typical GRP per capita for a province in 2019 is roughly ¥ 50,000 -- ¥80,000, with the highest GRP per capita (¥164,653) in Beijing and the lowest GRP per capita (¥32,995) in Gansu. Now let's look at the side-by-side boxplots of region and leading sector respectively.



For the region effect, the regions fall into three groups: central, east and west. We notice that both central and west lag behind the east, as east is far ahead of them. They have different variabilities in GRP per capita, with central and west having less variability and east having much higher variability. So, we roughly conclude that the non-constant variance exists.



For the leading sector effect, the finance sector generally has the highest GRP per capita. m&c sector ranks the second, and the service sector has the lowest GRP per

capita. Variability of the service sector is lower than that of m&c and finance sectors. So, we roughly conclude that the non-constant variance exists.

Now I try to fit a two-way ANOVA model, but it ended up in failure as shown below.

* NOTE * Some of the requested means were removed from the model.

The following terms cannot be estimated and were removed:
Region*Leading Sector

And let's take a look at the tabulated statistics.

Rows: Region Columns: Leading Sector

	finance	m&c	service	All
Central	0	2	6	8
East	4	5	2	11
West	0	1	11	12
All	4	8	19	31

Cell Contents
Count

In order to solve this problem of some never-occurring combinations, there are three possible solutions. 1) Fit the model with only the main effects -- Not this option, because I'm interested in exploring the interaction effects between regions and leading sectors. 2) Fit the interaction manually using indicator or effect coding variables, and determine the appropriate partial F-test by hand -- Not this option, because it's somehow troublesome. 3) Change our data to omit the "holes" (0s) – I will try this option. In this dataset, the problem mainly lies in the finance sector -- neither central or west has finance sectors. The finance sector only appears in the east. To address this, I will remove the finance sector for now. After removing this sector, we noticed that there are still (31-4=27) provinces remaining in the sample, which is a reasonable proportion of 87.10% (27/31x100%) of the original data. Here is a two-way ANOVA for 3 regions (east, central and west) and 2 leading sectors (m&c and service sectors).

General Linear Model: GRP Per Capita versus Region, Leading Sector

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
Region	Fixed	3	Central, East, West
Leading Sector	Fixed	2	m&c, service

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Region	2	655342402	327671201	1.32	0.288
Leading Sector	1	2045921901	2045921901	8.24	0.009
Region*Leading Sector	2	105097234	52548617	0.21	0.811
Error	21	5213018169	248238960		
Total	26	10766578271			

Model Summary

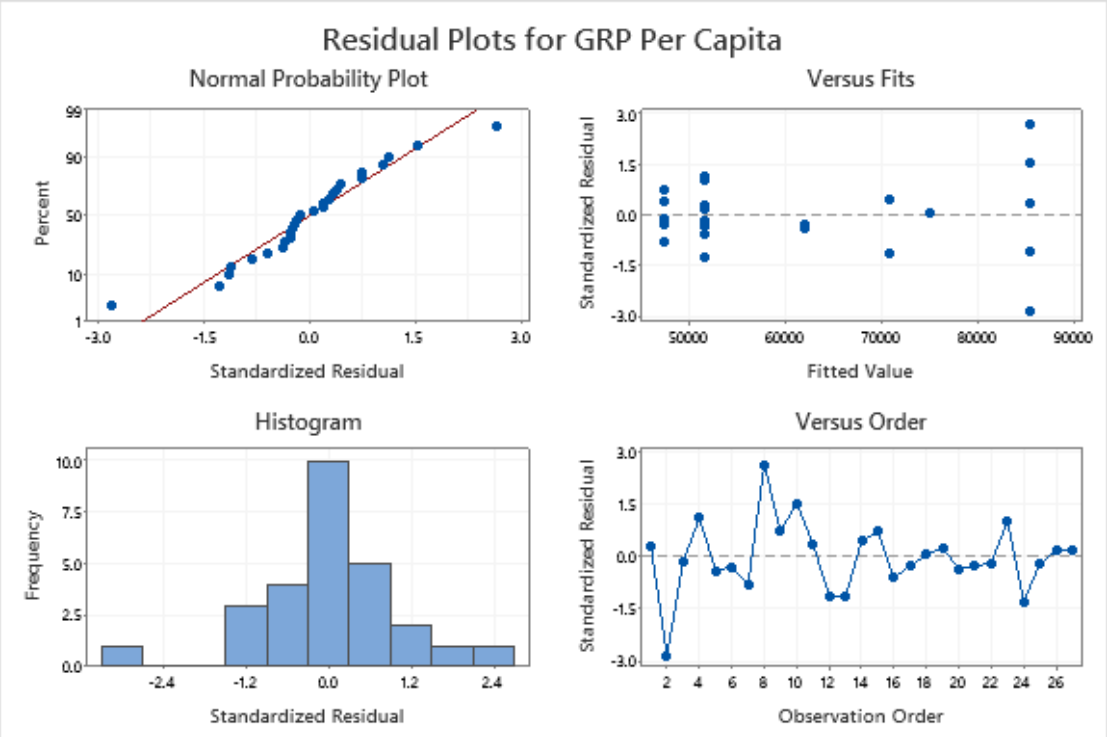
S	R-sq	R-sq(adj)	R-sq(pred)
15755.6	51.58%	40.05%	*

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	64596	4117	15.69	0.000	
Region					
Central	-6986	5544	-1.26	0.221	2.40
East	7929	5606	1.41	0.172	2.29
Leading Sector					
m&c	11818	4117	2.87	0.009	1.54
Region*Leading Sector					
Central m&c	-3159	5544	-0.57	0.575	2.31
East m&c	2803	5606	0.50	0.622	1.61

Regression Equation

GRP Per Capita = 64596 - 6986 Region_Central + 7929 Region_East - 942 Region_West
+ 11818 Leading Sector_m&c - 11818 Leading Sector_service
- 3159 Region*Leading Sector_Central m&c
+ 3159 Region*Leading Sector_Central service
+ 2803 Region*Leading Sector_East m&c - 2803 Region*Leading Sector_East service + 356 Region*Leading Sector_West m&c
- 356 Region*Leading Sector_West service



Let's first take a look at the residual plots. We notice that for the normal plot, it seems to be roughly on a straight line, which tested the normality of the errors. However, the residuals versus fitted values plot indicates obvious nonconstant variance. We can infer that the province Jiangsu and Hebei tend to be outliers. To check this, I output the standardized residuals, leverages and Cook's distance as below.

Province	**SRES_1**	**HI_1**	**COOK_1**
Tianjin	0.2066	0.2	0.001778
Hebei	-2.90689	0.2	0.352084
Shanxi	-0.23656	0.16667	0.001865
Inner Mongolia	1.08985	0.09091	0.019796
Liaoning	-0.07508	0.5	0.00094

Jilin	-0.38075	0.16667	0.004832
Heilongjiang	-0.9004	0.16667	0.027024
Jiangsu	2.50148	0.2	0.260724
Anhui	0.63414	0.16667	0.013404
Fujian	1.40641	0.2	0.082416
Jiangxi	0.27211	0.16667	0.002468
Shandong	-1.20759	0.2	0.060761
Henan	-0.9374	0.5	0.146454
Hubei	0.9374	0.5	0.146454
Hunan	0.61147	0.16667	0.012463
Guangxi	-0.56688	0.09091	0.005356
Hainan	0.07508	0.5	0.00094
Chongqing	*	1.00000	*
Sichuan	0.26575	0.09091	0.001177
Guizhou	-0.33595	0.09091	0.001881
Yunnan	-0.23537	0.09091	0.000923
Xizang	-0.1716	0.09091	0.000491
Shaanxi	1.00977	0.09091	0.016994
Gansu	-1.23049	0.09091	0.025235
Qinghai	-0.1937	0.09091	0.000625
Ningxia	0.18221	0.09091	0.000553
Xinjiang	0.1864	0.09091	0.000579

By checking standardized residuals, the absolute value for Hebei is |-2.90689| > 2.5 and the value for Jiangsu is 2.50148 > 2.5, so we can say both Hebei and Jiangsu are pretty unusual and could be outliers. Now we moved on to check leverages (though it's not really meaningful in ANOVA data), and the threshold for reference

here is 2.5(p+1/n)=2.5(2+1)/27=0.2778. The meaning of the leverage points here is how unbalanced our data are. Provinces having these leverage points greater than 0.2778 are: Liaoning(0.5), Henan(0.5), Hubei(0.5), Hainan(0.5) and Chongqing(1). We can observe that their unusualness is since their regions “east” and “central” are fewer than “west” in quantity. We do not need to take them out, or else we won’t be able to fit an interaction effect. Lastly we checked Cook's distance, and it turned out to be no values greater than 1, which means the effects of Jiangsu and Hebei as outliers are not dramatically but moderately influential. After careful observation, I found that for Hainan and Liaoning as leverage points, it’s because their GDP per capita is particularly lower than other provinces in the same region category of “east”. For Henan and Hubei as leverage points, it’s because their GDP per capita is particularly higher than other provinces in the same region “central”. For Chongqing as a leverage point, it’s because it is the only province falling into the category of “west” and “m&c”. Let’s ignore the outliers for now and take a look at the interaction effect. The P-value of the interaction term is 0.811>0.05, which is way far from being statistically significant. In other words, this is saying that we do not need the interaction term. But now let’s first deal with the outliers by omitting them from the dataset.

General Linear Model: GRP Per Capita versus Region, Leading sector

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
Region	Fixed	3	Central, East, West
Leading sector	Fixed	2	m&c, service

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Region	2	674174871	337087436	2.81	0.085
Leading sector	1	2046274384	2046274384	17.04	0.001
Region*Leading sector	2	128053764	64026882	0.53	0.595
Error	19	2281378296	120072542		
Total	24	6587306434			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
10957.8	65.37%	56.25%	*

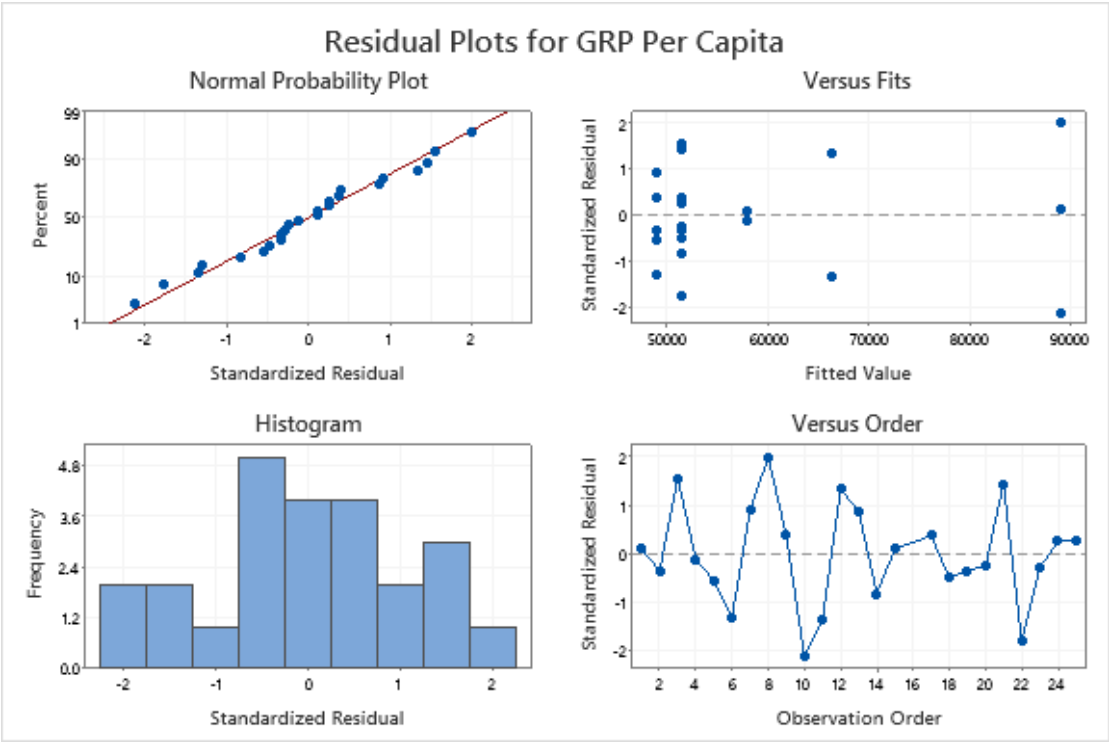
Coefficients

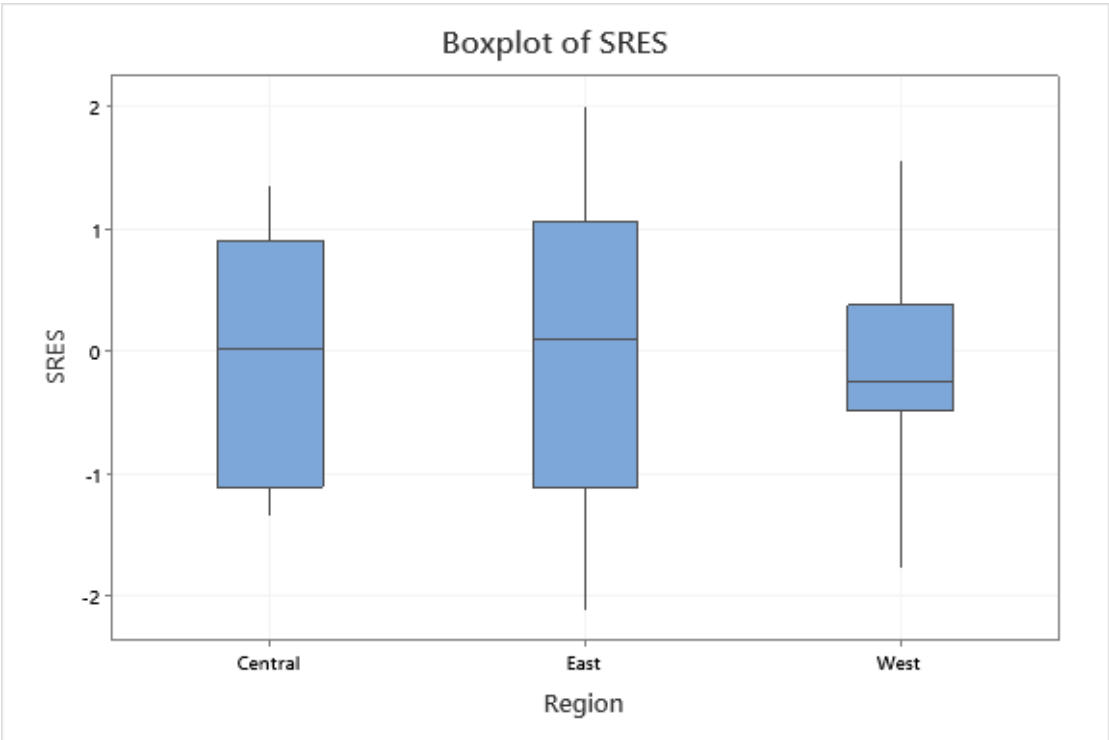
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	64914	2940	22.08	0.000	
Region					
Central	-7304	3913	-1.87	0.077	2.47
East	8564	4121	2.08	0.051	2.13
Leading sector					
m&c	12135	2940	4.13	0.001	1.31
Region*Leading sector					
Central m&c	-3477	3913	-0.89	0.385	2.37
East m&c	3438	4121	0.83	0.414	1.72

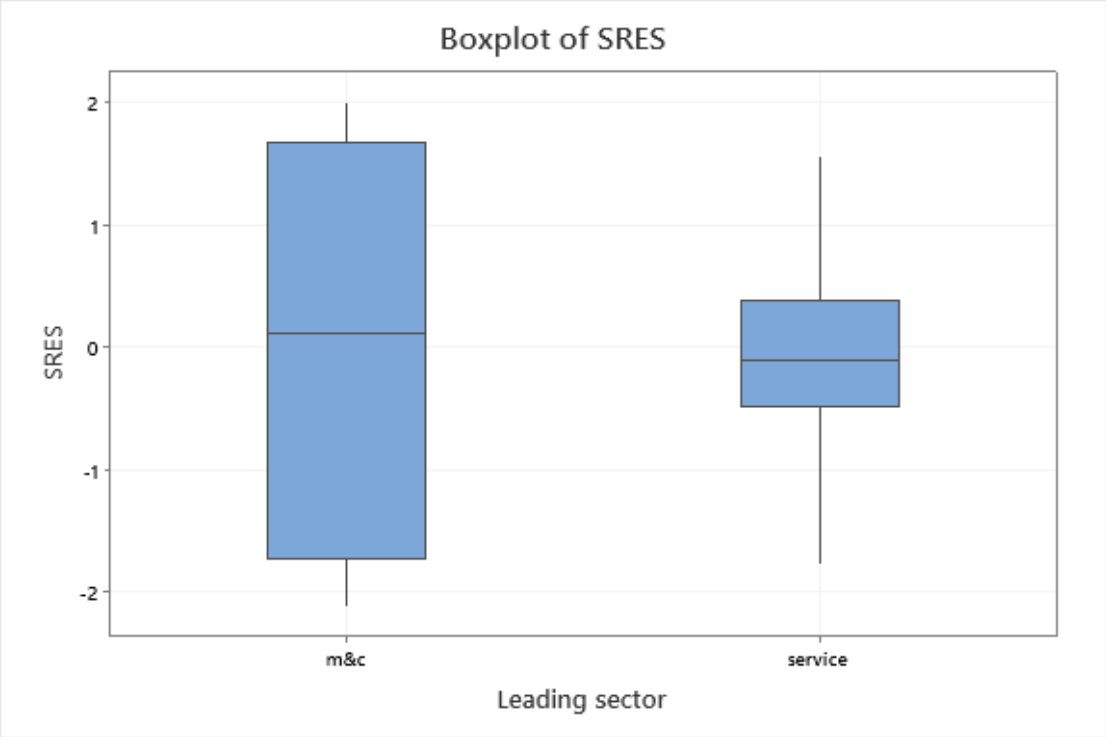
Regression Equation

GRP Per Capita = 64914 - 7304 Region_Central + 8564 Region_East - 1260 Region_West
+ 12135 Leading sector_m&c - 12135 Leading sector_service
- 3477 Region*Leading sector_Central m&c
+ 3477 Region*Leading sector_Central service
+ 3438 Region*Leading sector_East m&c - 3438 Region*Leading sector_East service + 39 Region*Leading sector_West m&c - 39 Region*Leading sector_West service

We can see from the large P-value (0.595) of the interaction term that the interaction effect is not statistically significant. Let's also look at the residual plots as well as side-by-side boxplots on residuals of region and leading sector respectively, to see if the assumptions of the regression seem reasonable.


Residual Plots for GRP Per Capita
Normal Probability Plot
Percent
Standardized Residual
Versus Fits
Standardized Residual
Fitted Value
Histogram
Frequency
Standardized Residual
Versus Order
Standardized Residual
Observation Order


Boxplot of SRES
SRES
Central
East
West
Region



So we can see that there is non-constant variance related to both leading sectors and regions. Since the interaction effect is not statistically significant, we will omit this term and only focus on main effects. Note that I do not omit the outliers now, as I'm also wondering what the outliers may change after I omit the interaction effect. Also, I will put back the finance sector, as there's no reason to ignore it when we do not have the interaction effect. Here is a two-way ANOVA for 3 regions (east, central and west) and 3 leading sectors (finance, m&c and service sectors).

General Linear Model: GRP Per Capita versus Region, Leading Sector

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
Region	Fixed	3	Central, East, West
Leading Sector	Fixed	3	finance, m&c, service

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Region	2	663289431	331644715	0.96	0.396
Leading Sector	2	9009034231	4504517115	13.05	0.000
Error	26	8972837806	345109146		
Lack-of-Fit	2	105097234	52548617	0.14	0.868
Pure Error	24	8867740571	369489190		
Total	30	31912840265			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
18577.1	71.88%	67.56%	56.20%

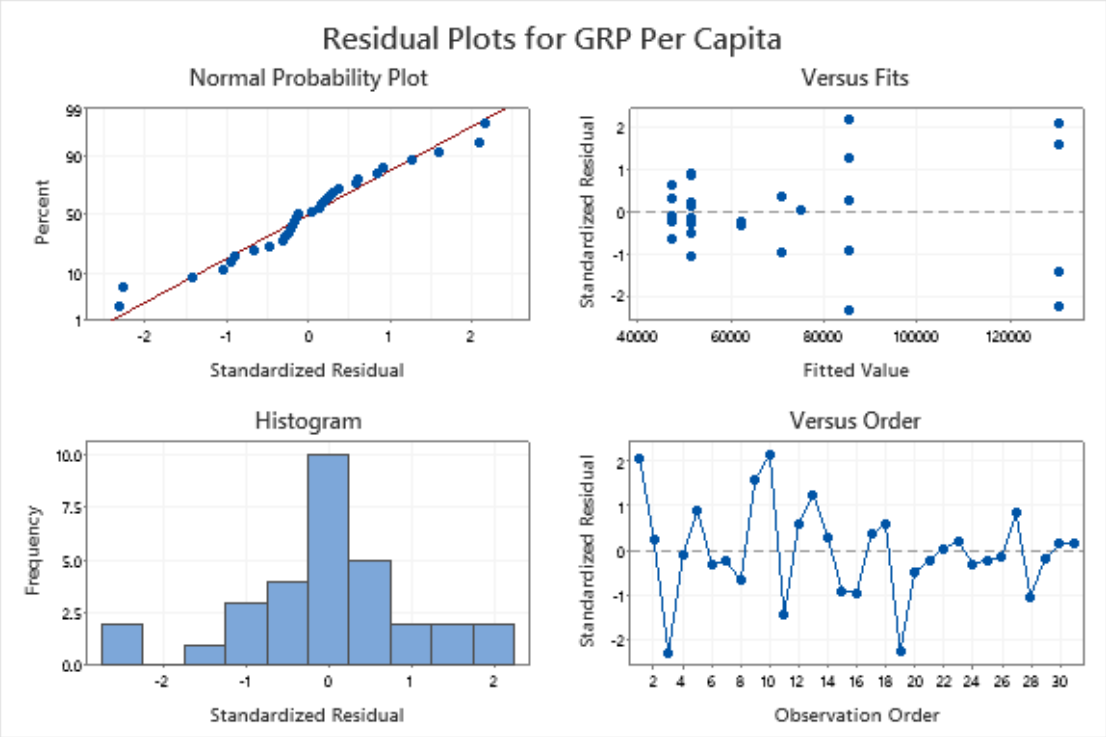
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	84419	4691	18.00	0.000	
Region					
Central	-6256	5360	-1.17	0.254	1.62
East	8380	6470	1.30	0.207	2.79
Leading Sector					
finance	38054	7880	4.83	0.000	2.83
m&c	-7315	5617	-1.30	0.204	2.11

Regression Equation

GRP Per Capita = 84419 - 6256 Region_Central + 8380 Region_East - 2124 Region_West + 38054 Leading Sector_finance - 7315 Leading Sector_m&c - 30739 Leading Sector_service

We notice that the P-value of the region is 0.396>0.05, meaning that this main effect is not significant. We also get this conclusion by looking at the partial-F test of the region predictor as its F-Value is as small as 0.96. This tells us that given the presence of the leading sector variable, there is not much significance of the region variable. Now we may consider only including one main effect, but before that, let's find whether there are outliers by looking at the residual plots.

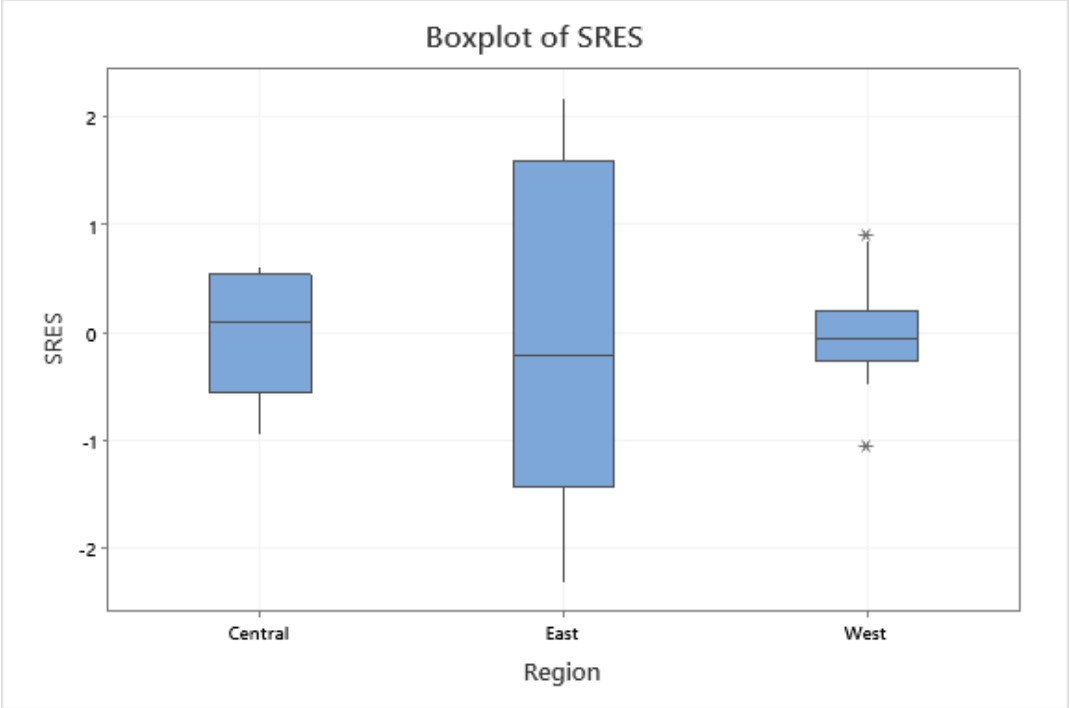


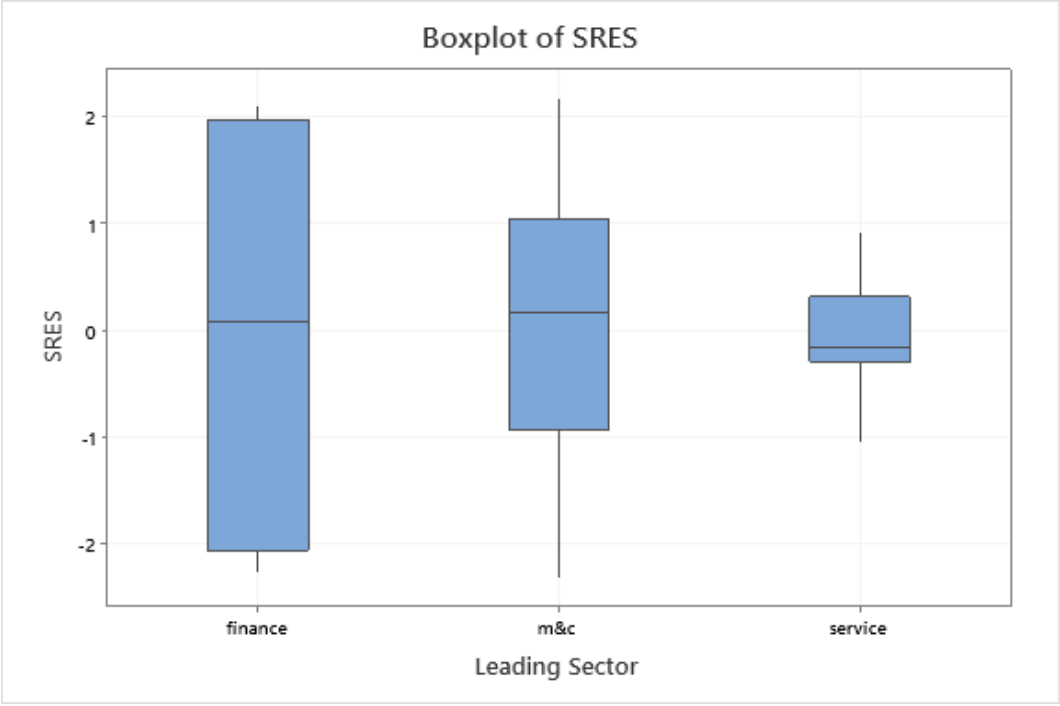
There are potential outliers and non-constant variance. So we look at the standardized residuals, leverages and Cook's distance to double check whether there are outliers.

Province	**SRES**	**HI**	**COOK**
Beijing	2.09532	0.25	0.29269
Tianjin	0.2693	0.164087	0.002847
Hebei	-2.31396	0.164087	0.21021
Shanxi	-0.10895	0.141254	0.000391
Inner Mongolia	0.91707	0.085139	0.015653
Liaoning	-0.31578	0.275542	0.007585
Jilin	-0.22943	0.141254	0.001732
Heilongjiang	-0.66358	0.141254	0.014486
Shanghai	1.59955	0.25	0.170571
Jiangsu	2.17336	0.164087	0.185441
Zhejiang	-1.43204	0.25	0.136716
Anhui	0.61849	0.141254	0.012584

Fujian	1.26478	0.164087	0.062802
Jiangxi	0.31602	0.141254	0.003285
Shandong	-0.90405	0.164087	0.032087
Henan	-0.94736	0.271285	0.066824
Hubei	0.36974	0.271285	0.010178
Hunan	0.59955	0.141254	0.011825
Guangdong	-2.26283	0.25	0.34136
Guangxi	-0.48359	0.085139	0.004353
Hainan	-0.20997	0.275542	0.003354
Chongqing	0.05458	0.301858	0.000258
Sichuan	0.22034	0.085139	0.000904
Guizhou	-0.28836	0.085139	0.001548
Yunnan	-0.20333	0.085139	0.000769
Xizang	-0.14941	0.085139	0.000415
Shaanxi	0.84937	0.085139	0.013428
Gansu	-1.04464	0.085139	0.020311
Qinghai	-0.1681	0.085139	0.000526
Ningxia	0.14971	0.085139	0.000417
Xinjiang	0.15326	0.085139	0.000437

By checking standardized residuals, no value is greater than 2.5, indicating there are no potential outliers. For leverages, the threshold for reference here is 2.5(p+1/n)=2.5(2+1)/31=0.2419. Though there are some leverage points (marked in red) greater than 0.2419, notice that in this ANOVA data context it's not meaningful for us to make further interpretation, as it just means how unbalanced our data are, which is pretty common for observational data. For the Cook's distance, no value is greater than 1, meaning there's no leverage points having dramatic influence. Now let's check the side-by-side boxplots of the residuals on "region" and "leading sector" respectively to make sure if there's non-constant variance.





It’s clear that there is non-constant variance, and we can formalize this conclusion by looking at Levene’s test.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Region	2	0.5807	0.2903	1.18	0.325
Leading Sector	2	2.2442	1.1221	4.54	0.020
Error	26	6.4226	0.2470		
Lack-of-Fit	2	1.3217	0.6609	3.11	0.063
Pure Error	24	5.1009	0.2125		
Total	30	14.7455			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.497016	56.44%	49.74%	34.27%

The Levene's test shows that there is moderately strong evidence saying that non-constant variance is related to the leading sector as its P-value is relatively small 0.01<0.020<0.05, so we can reject the null hypothesis that variances are equal. For the region, we do not see non-constant variance as its P-value (0.325) is big, meaning we don't have enough evidence to reject the null hypothesis that they are the same. Therefore, there could be non-constant variance related to the leading sector. In order to handle this heteroscedasticity, we use weighted least squares. First, we determine the weights by using the standard deviations of residuals. Here's the output of the test for equal variances.

Test for Equal Variances: SRES versus Region, Leading Sector

95% Bonferroni Confidence Intervals for Standard Deviations

Region	Leading Sector	N	StDev	CI
Central	m&c	2	0.93133	(*, *)
Central	service	6	0.50981	(0.209536, 2.2138)
East	finance	4	2.16949	(0.456811, 30.2652)
East	m&c	5	1.76915	(0.504344, 13.1383)
East	service	2	0.07482	(*, *)
West	m&c	1	*	(*, *)
West	service	11	0.56186	(0.292492, 1.4198)

Individual confidence level = 99.1667%

Now we calculate the weights of each combination by first taking the squared standard deviation to get the variance and then taking the inverse of it. Note that there's only one data (Chongqing) in the combination of west and m&c, and I choose to omit it because I don't find it significant to include this special data in our analysis. Also, I will omit the two data under "east" and "service", because there are only two data in this combination and the standard deviation is extremely small, meaning that the GRP per capita of these two provinces are very close to each other compared to other provinces. We find that these two provinces are Liaoning (57067) and Hainan (58740). So I calculated their weights after omitting Chongqing, Liaoning and Hainan. Now let's fit a weighted least square model of the two-way ANOVA (3 regions: east,west and central; 3 leading sectors: m&c, service and finance) on the remaining 28 observations.

General Linear Model: GRP Per Capita versus Region, Leading Sector

Method

Factor coding	(-1, 0, +1)
Weights	weight

Factor Information

Factor	Type	Levels	Values
Region	Fixed	3	Central, East, West
Leading Sector	Fixed	3	finance, m&c, service

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Region	2	499690451	249845225	0.85	0.441
Leading Sector	2	1679480540	839740270	2.85	0.078
Error	23	6776527923	294631649		
Total	27	14596211768			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
17164.8	53.57%	45.50%	30.92%

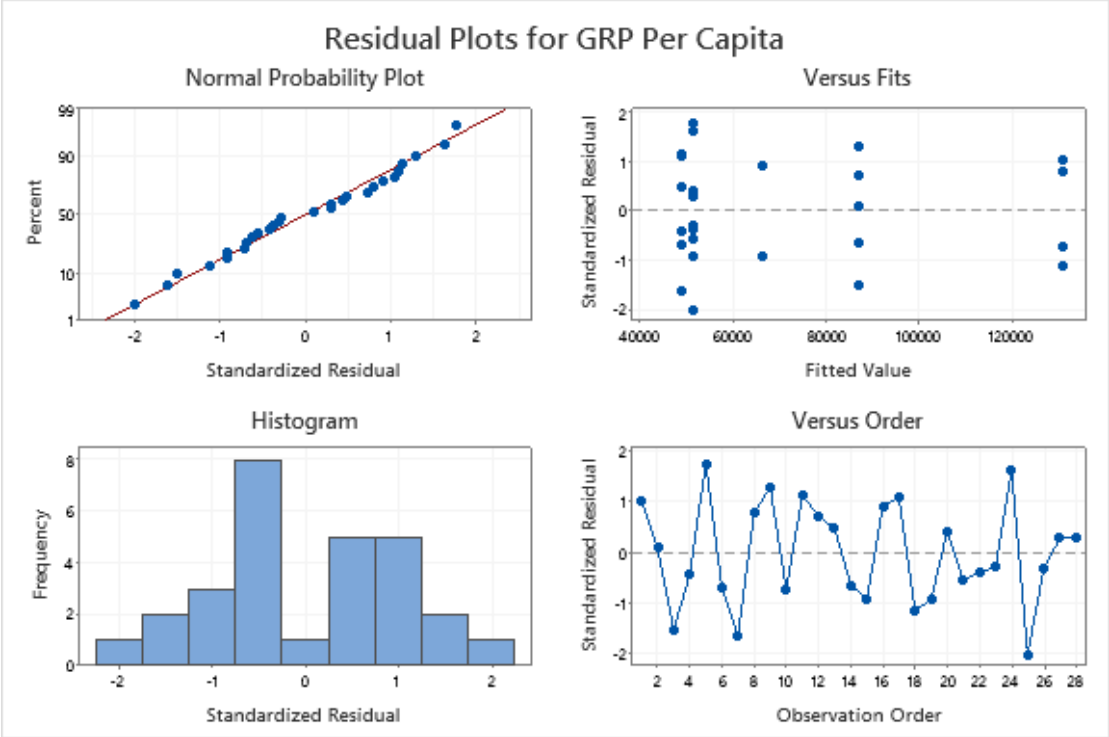
Coefficients

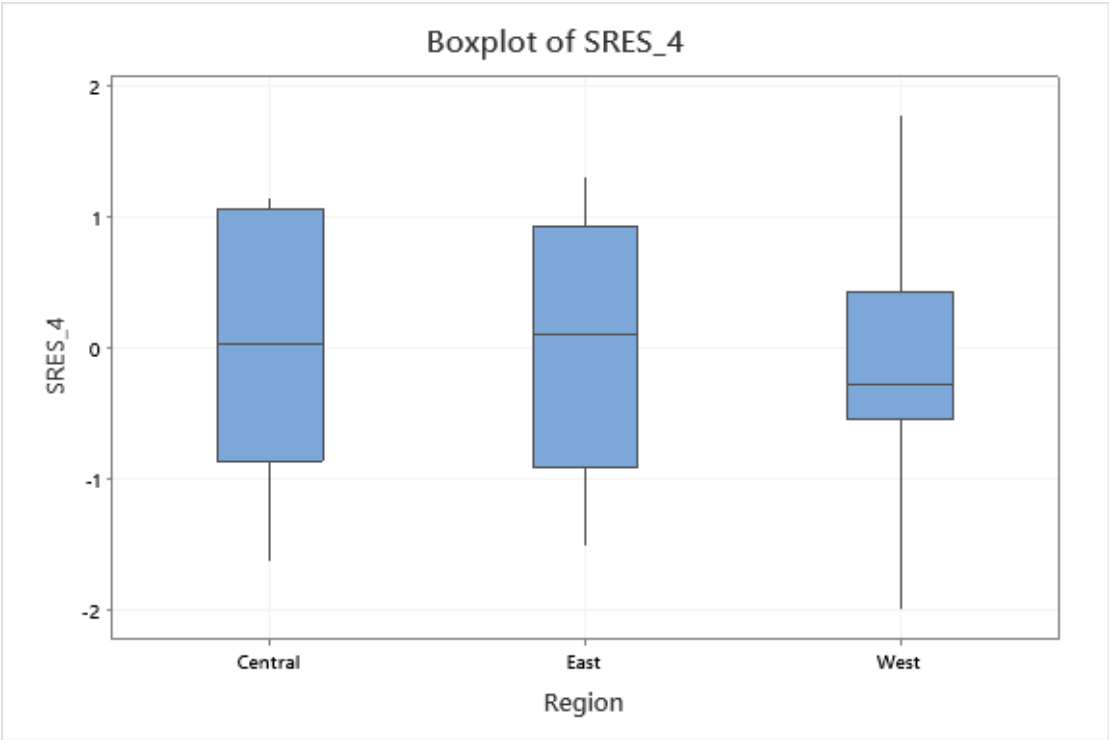
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	82867	7359	11.26	0.000	
Region					
Central	-7802	6088	-1.28	0.213	7.36
East	13076	11880	1.10	0.282	9.79
Leading Sector					
finance	34910	15918	2.19	0.039	5.81
m&c	-8796	8665	-1.02	0.321	3.88

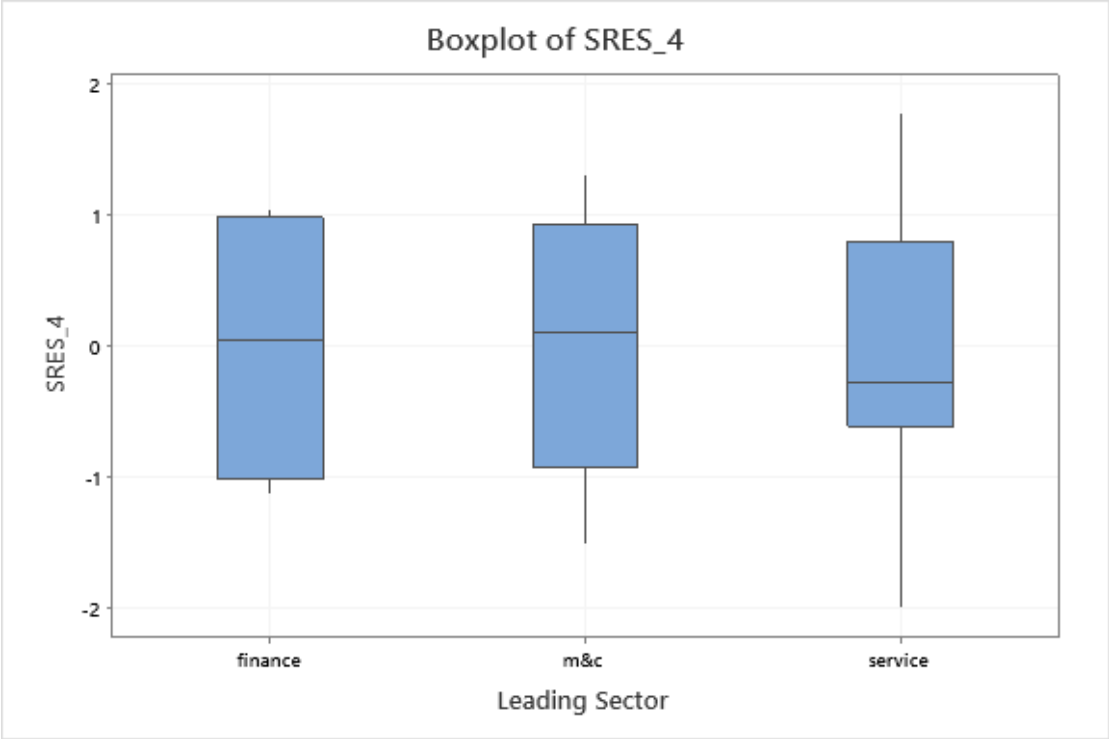
Regression Equation

GRP Per Capita = 82867 - 7802 Region_Central + 13076 Region_East - 5274 Region_West
+ 34910 Leading Sector_finance - 8796 Leading Sector_m&c
- 26114 Leading Sector_service

We see that the P-values of both “region” and “leading sector” are not small, meaning that holding either variable fixed, the presence of the other variable does not add any significance. We admit this result and move on to see if non-constant variance still occurs by taking a look at the residual plots, side-by-side boxplots of each predictor and the standardized residuals and its Cook’s distance.







Province	**SRES_4**	**HI_4**	**COOK_4**
Beijing	1.0392	0.25	0.071995
Tianjin	0.10727	0.2	0.000575
Hebei	-1.50938	0.2	0.113912

Shanxi	-0.42605	0.166667	0.007261
Inner Mongolia	1.76985	0.090909	0.062647
Jilin	-0.68576	0.166667	0.018811
Heilongjiang	-1.62167	0.166667	0.105192
Shanghai	0.79332	0.25	0.041957
Jiangsu	1.29887	0.2	0.084354
Zhejiang	-0.71024	0.25	0.033629
Anhui	1.14211	0.166667	0.052177
Fujian	0.73027	0.2	0.026664
Jiangxi	0.49008	0.166667	0.009607
Shandong	-0.62703	0.2	0.019658
Henan	-0.92272	0.5	0.170283
Hubei	0.92272	0.5	0.170283
Hunan	1.10129	0.166667	0.048513
Guangdong	-1.12228	0.25	0.083967
Guangxi	-0.92057	0.090909	0.016949
Sichuan	0.43156	0.090909	0.003725
Guizhou	-0.54556	0.090909	0.005953
Yunnan	-0.38222	0.090909	0.002922
Xizang	-0.27866	0.090909	0.001553
Shaanxi	1.6398	0.090909	0.053779
Gansu	-1.99822	0.090909	0.079858
Qinghai	-0.31455	0.090909	0.001979
Ningxia	0.29589	0.090909	0.001751
Xinjiang	0.3027	0.090909	0.001833

We see that there's no outliers as no standardized residual is greater than 2.5 and Cook's distance greater than 1. Also, non-constant variance is surely eased, which has seen a great improvement from before, though there's still variability in the west region. Now we formalize our conclusion by using Levene's test.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Region	2	0.1537	0.076844	0.10	0.909
Leading Sector	2	0.0024	0.001188	0.00	0.999
Error	23	18.5179	0.805125		
Total	27	18.6827			

The Levene's test shows that there is no non-constant variance issue related to both region and leading sector, as their P-values are very big (0.999>0.05, 0.909>0.05). So we are pretty confident to say the non-constant variance has been removed from our current model. Now we want more interpretations of this model (though we know this model is not perfect as we discussed before). First, let's see the prediction on all of the five combinations (central-m&c, central-service, east-finance, east-m&c, west-service).

General Linear Model Information

Terms
Region Leading Sector

Settings

Variable	Setting
Region	Central
Leading Sector	service

Prediction

Fit	SE Fit	95% CI	95% PI
48951.3	3571.36	(41563.4, 56339.3)	(29404.7, 68497.9)

Weight = 3.85

Settings

Variable	Setting
Region	Central
Leading Sector	m&c

Prediction

Fit	SE Fit	95% CI	95% PI
66268.5	11318.2	(42855.1, 89681.9)	(25715.3, 106822)

Weight = 1.15

Settings

Variable	Setting
Region	East
Leading Sector	finance

Prediction

Fit	SE Fit	95% CI	95% PI
130853	18728.4	(92110.4, 169596)	(44221.9, 217484)

Weight = 0.21

Settings

Variable	Setting
Region	East
Leading Sector	m&c

Prediction

Fit	SE Fit	95% CI	95% PI
87146.6	13570.0	(59074.9, 115218)	(18385.3, 155908)

Weight = 0.32

Settings

Variable	Setting
Region	West
Leading Sector	service

Prediction

Fit	SE Fit	95% CI	95% PI
51479.8	2925.31	(45428.4, 57531.3)	(30517.0, 72442.7)

Weight = 3.13

We see that for the central region, the m&c sector tends to have a ¥17317.2 (66268.5-48951.3) higher GRP per capita in 2019 then the service sector. We see that for the east region, the finance sector tends to have a ¥43706.4 (130853-87146.6) higher GRP per capita in 2019 then the m&c sector. For the service sector, the west region tends to have a ¥2528.5 (51479.8-48951.3) higher GRP per capita in 2019 than the east central region. For the prediction intervals, all provide us with a lot of information except that for the east-finance combinations, as the upper bound is as high as 217484, which is even greater than our biggest data (164563). Therefore, this prediction interval does not provide us with useful information. Now let's move on to choose the best subset model by running the general linear model three times and displaying the expanded tables as we want to see AICc. There are three possible models in total: 1) only region as the predictor, 2) only leading sector as the predictor, and 3) region and leading sector as the predictors. I will not include the interaction term as we've verified before that the interaction term does not add any significance.

General Linear Model: GRP Per Capita versus Region

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
18391.3	42.07%	37.43%	1.05704E+10	27.58%	626.95	630.54

General Linear Model: GRP Per Capita versus Leading Sector

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
17060.2	50.15%	46.16%	9160313308	37.24%	622.74	626.33

General Linear Model: GRP Per Capita versus Region, Leading Sector

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
17164.8	53.57%	45.50%	1.00837E+10	30.92%	627.01	631.00

As we compare the AICcs of all the three models, we see the second model with the leading sector as the only predictor that has the smallest AICc, meaning that this model is better than the other two, though the AICc of these three models do not differ a lot. So now we can try to fit a one-way ANOVA with "leading sector" as the only predictor and GRP per capita as the response variable.

General Linear Model: GRP Per Capita versus Leading Sector

Method

Factor coding	(-1, 0, +1)
Weights	weight

Factor Information

Factor	Type	Levels	Values
Leading Sector	Fixed	3	finance, m&c, service

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Leading Sector	2	7319993394	3659996697	12.58	0.000
Error	25	7276218374	291048735		
Total	27	14596211768			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
17060.2	50.15%	46.16%	37.24%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	85384	6881	12.41	0.000	
Leading Sector					
finance	45469	12761	3.56	0.002	3.78
m&c	-10550	8499	-1.24	0.226	3.78

Regression Equation

GRP Per Capita = 85384 + 45469 Leading Sector_finance - 10550 Leading Sector_m&c - 34919 Leading Sector_service

The P-value of “leading sector” is 0.000<0.001, which means there is a statistically significant relationship between GRP per capita and leading sector. We also want to look at the multiple comparisons, trying to see if all of the leading sectors are distinct.

Tukey Pairwise Comparisons: Leading Sector

Grouping Information Using the Tukey Method and 95% Confidence

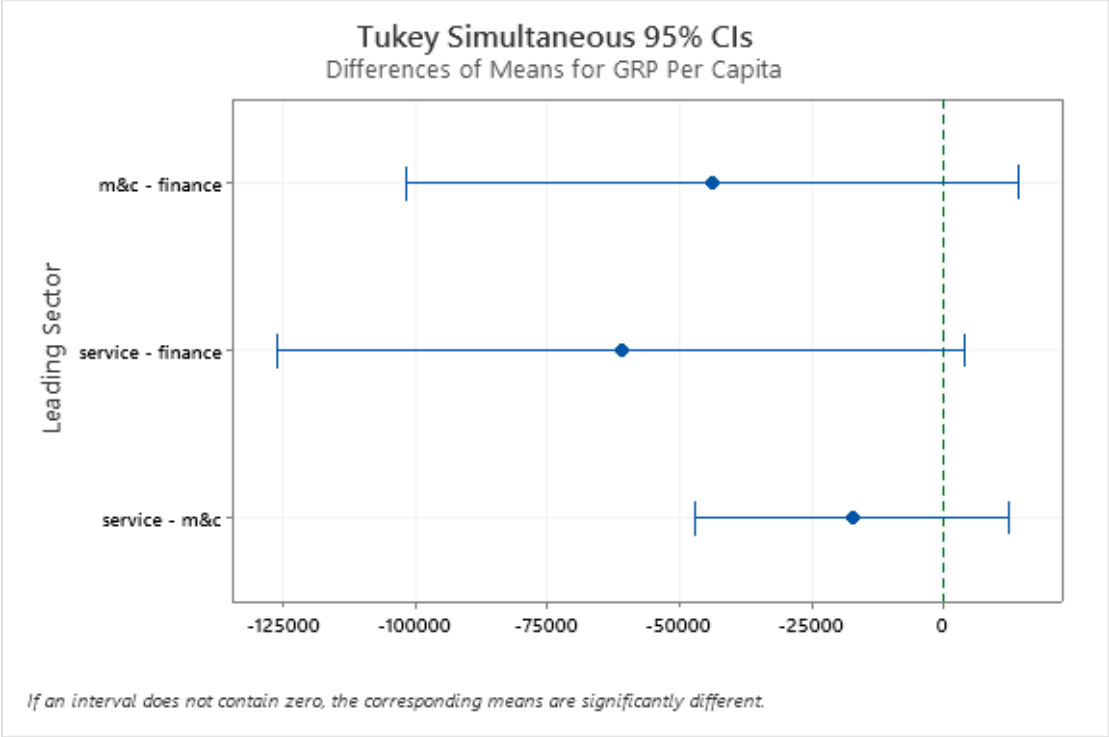
Leading Sector	N	Mean	Grouping
finance	4	117777	A
m&c	7	74071	A
service	17	56754	A

Means that do not share a letter are significantly different.

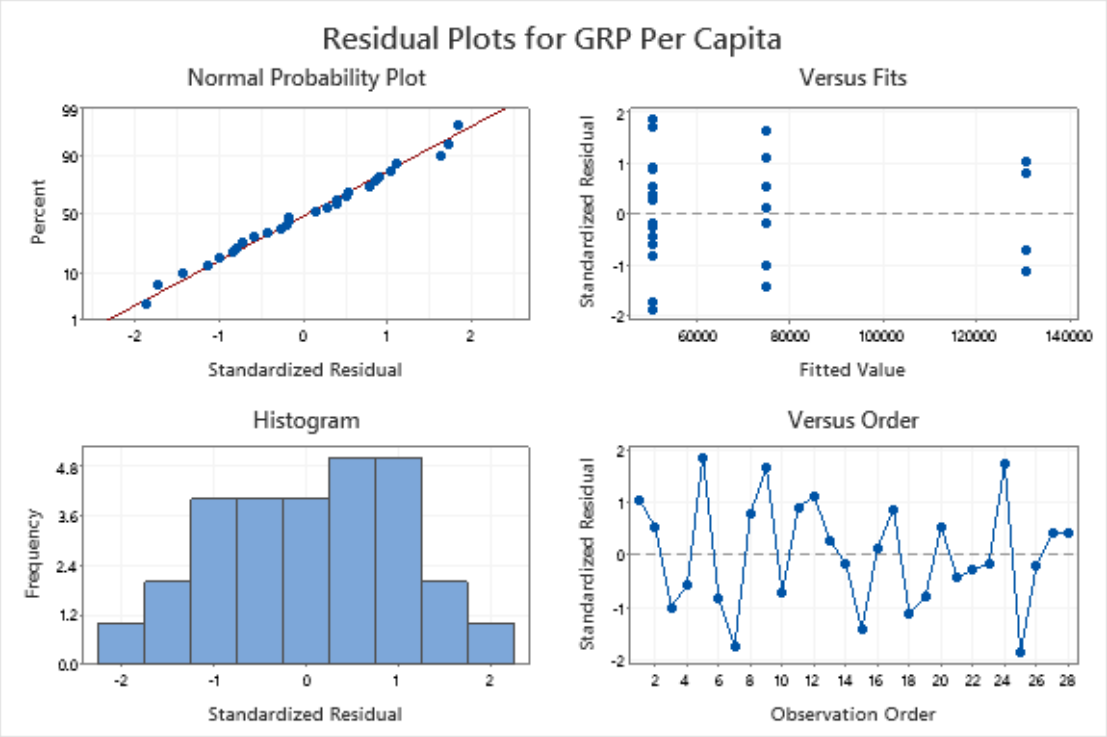
Tukey Simultaneous Tests for Differences of Means

Difference of Leading Sector Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
m&c - finance	-43706	23128	(-101599, 14186)	-1.89	0.164
service - finance	-61024	25995	(-126094, 4047)	-2.35	0.069
service - m&c	-17317	11868	(-47025, 12391)	-1.46	0.329

Individual confidence level = 98.01%

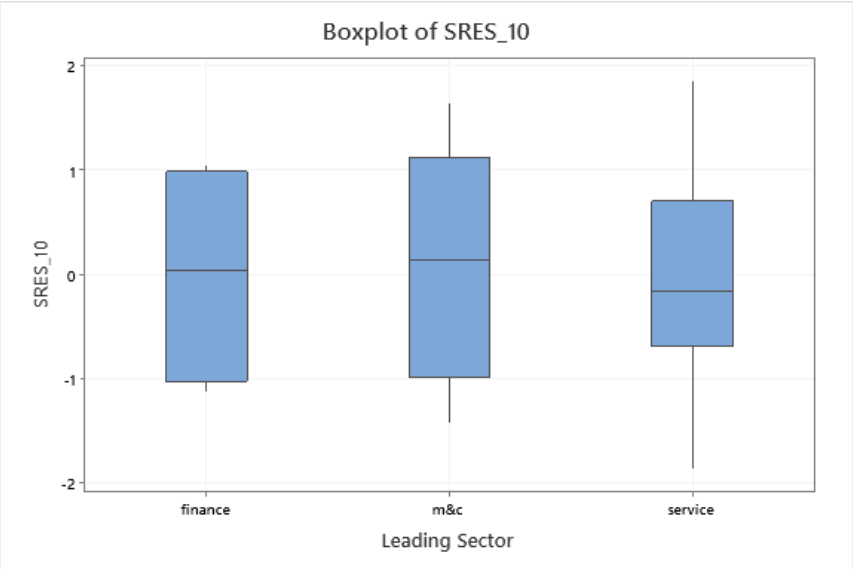


As we can see from the Tukey pairwise comparisons, all leading sectors are labeled with Grouping A, meaning that these sectors are not statistically different from each other. The mean of the finance sector is the biggest of the three sectors. From the graph of Tukey Simultaneous 95% CIs, all intervals contain zero, meaning their corresponding means are not significantly different from one another. Though this is not a perfect result that we anticipate seeing, I still frankly report what I observed. Now let's look at its residual plots and the Levene's test to see if there's non-constant variance.



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Leading Sector	2	0.0089	0.004468	0.01	0.995
Error	25	20.4994	0.819975		
Total	27	20.5083			



We can see from the residual plots and the large P-value shown by Levene's test that the non-constant variance has been addressed well. We also get this conclusion by observing that the difference in variability for different sectors is small from the side-by-side boxplots. There are no outliers as we can see from the standardized

residuals (no values greater than 2.5) and Cook's distance (no values greater than 1) below.

Province	**SRES_10**	**HI_10**	**COOK_10**
Beijing	1.04557	0.25	0.12147
Tianjin	0.52688	0.082051	0.008271
Hebei	-0.9916	0.082051	0.029297
Shanxi	-0.58528	0.066922	0.008189
Inner Mongolia	1.85427	0.054406	0.065943
Jilin	-0.83222	0.066922	0.016558
Heilongjiang	-1.72212	0.066922	0.070901
Shanghai	0.79818	0.25	0.070789
Jiangsu	1.64612	0.082051	0.080737
Zhejiang	-0.71459	0.25	0.056738
Anhui	0.90579	0.066922	0.019615
Fujian	1.11205	0.082051	0.036846
Jiangxi	0.28581	0.066922	0.001953
Shandong	-0.16283	0.082051	0.00079
Henan	-1.42295	0.294872	0.282241
Hubei	0.14059	0.294872	0.002755
Hunan	0.86697	0.066922	0.01797
Guangdong	-1.12916	0.25	0.141668
Guangxi	-0.79989	0.054406	0.012271
Sichuan	0.53401	0.054406	0.005469
Guizhou	-0.42994	0.054406	0.003545
Yunnan	-0.2688	0.054406	0.001386
Xizang	-0.16664	0.054406	0.000533
Shaanxi	1.72598	0.054406	0.057134
Gansu	-1.86303	0.054406	0.066567
Qinghai	-0.20204	0.054406	0.000783
Ningxia	0.40018	0.054406	0.003071
Xinjiang	0.40689	0.054406	0.003175

Now we want the prediction intervals for different groups, and we included their weights by using the inverse of the squared standard deviations.

95% Bonferroni Confidence Intervals for Standard Deviations

Leading Sector	N	StDev	CI
finance	4	0.11517	(0.033206, 0.99492)
m&c	7	0.60547	(0.286912, 1.94181)
service	17	1.24292	(0.836490, 2.14954)

Individual confidence level = 98.3333%

Prediction for GRP Per Capita

General Linear Model Information

Terms
Leading Sector

Settings

Variable	Setting
Leading Sector	finance

Prediction

Fit	SE Fit	95% CI	95% PI	
130853	18614.2	(92516.4, 169190)	(92303.5, 169403)	XX

Weight = 75.39
XX denotes an extremely unusual point relative to predictor levels used to fit the model.

Settings

Variable	Setting
Leading Sector	m&c

Prediction

Fit	SE Fit	95% CI	95% PI	
74833.9	8638.74	(57042.1, 92625.7)	(47107.3, 102560)	XX

Weight = 2.73
XX denotes an extremely unusual point relative to predictor levels used to fit the model.

Settings

Variable	Setting
Leading Sector	service

Prediction

Fit	SE Fit	95% CI	95% PI
50464.6	2249.24	(45832.2, 55097.0)	(6638.16, 94291.0)

Weight = 0.65

We see that the prediction interval for the service sector is the narrowest, which can provide us with the most accurate and useful information. So until now, we are done with this one-way ANOVA and tested the constant variance, also acknowledging the difference in GRP per capita among all three leading sectors are not really significant.

· Conclusion / Lessons Learnt from this Report

This analysis did not go as smoothly as my previous projects, as firstly we found that after omitting the interaction effect and dealing with non-constant variance by WLS, the two main effects are not significant with the existence of one another. Acknowledging this, we chose the best subset model by displaying the expanded table under the general linear model and comparing different models' AICc. Finally, we chose the best subset model -- which turned out to be one-way ANOVA with only the leading sector as the predictor. We verified that there's no non-constant variance, yet this model shows that the difference between each leading sector is not significant. I will conclude that these three divisions of the leading sector variable may not be a good decision. Perhaps I shouldn't have consolidated some categories at the beginning such as the manufacturing and construction sectors, or I should have taken some more representative and differentiated categories. In all, no matter what my conclusion is, I gained a lot from trying different options, exploring both two-way and one-way ANOVA, as well as the interpretation process.