Project Name: Analysis of Factors Affecting Fiscal Revenue

Professor: Nizar Zaarour
Student Name: Xiaofan Wang (NUID: 001302984),
Cheng Peng (NUID: 001497625)
Course Name: Statistical Methods in Engineering
Course Number: IE7280
Course Time: Thursday 6.00-9.15pm

Analysis of Factors Affecting Fiscal Revenue

Group 2: Xiaofan Wang (NUID: 001302984), Cheng Peng (NUID: 001497625)

1. Introduction

1.1. Background

Local fiscal revenue is a comprehensive reflection for the level of regional national economy, and it is also the basis for macroeconomic control for the governments that are dominated by market economy. The analysis of factors affecting fiscal revenue and the prediction of its outcomes have important guiding significance for regional economic development, tax management systems, and fiscal policies; Meanwhile, it also has a significant impact on a steady and rapid development of the national economy.

1.2. Feature Selection and Interpretation

According to the existing relevant research, combined with economic theories' interpretation of fiscal revenue, as well as the official principles and definition on National Bureau of Statistics of China, this paper has selected 7 main features as explanatory variables of fiscal revenue (y), their specific meanings can be interpreted as following:

(1) Total retail sales of social consumer goods (x1)
Reflecting the overall consumption level of society. Generally, the expansion of people's consumption demand will cause obvious changes on taxation and other aspects of the economic system, thereby driving the growth of fiscal revenue.

(2) Gross Domestic Product (GDP, x2)
Gross Domestic Product (GDP) is the monetary value of all finished goods and services made within a country during a specific period. The sum of the added value of various industries reflects the level of economic development of a region. GDP provides an economic snapshot of a country, used to estimate the size of an economy and growth rate. GDP can be calculated in three ways, using expenditures, production, or incomes. Generally, the more developed the region's economy, the higher GDP, and the more fiscal revenue.

(3) Total population at the end of each year (x3)
The relationship between population and economy has mutual interaction. With the remaining factors unchanged, the larger the total population, the lower the per capita fiscal income.

(4) Total social investment in fixed assets (x4)
It is the main means for the reproduction of social fixed assets. It can adjust the economic structure and enhance economic strength, and affect fiscal revenue to some certain extent.

(5) Consumer Price Index (CPI, x5)

The CPI is a statistical estimate constructed using the prices of a sample of representative items whose prices are collected periodically. Sub-indices and sub-sub-indices are computed for different categories and sub-categories of goods and services, being combined to produce the overall index with weights reflecting their shares in the total of the consumer expenditures covered by the index. It is one of several price indices calculated by most national statistical agencies. The annual percentage change in a CPI is used as a measure of inflation. A CPI can be used to index (i.e. adjust for the effect of inflation) the real value of wages, salaries, and pensions; to regulate prices; and to deflate monetary magnitudes to show changes in real values.

(6) Tax (x6)

Tax is a compulsory financial charge or some other type of levy imposed upon a taxpayer by a governmental organization in order to fund various public expenditures. Taxation is the most important form and source of revenue for the state (government) public finances. The essence of taxation is a special distribution relationship formed by the state in order to meet the public needs of society, relying on public power, participating in the distribution of national income, and compulsively obtaining fiscal revenue in accordance with the standards and procedures prescribed by law. It reflects a specific distribution relationship between the state and taxpayers in the collection and distribution of tax benefits under a certain social system.

(7) Number of employees

Employed persons refer to persons who are 18 years of age and above, who are engaged in certain social labor and have obtained certain labor remuneration or operating income.

1.3. Dataset

(1) Data Source: In this project, the dataset has been collected from the website of National Bureau of Statistics of China (http://www.stats.gov.cn/).

(2) Data size: 49 rows * 8 columns

(3) The range of data: Data records come from 1970 to 2018, about 49 years in total.

2. Data Presentation

Table 2 Factors affecting fiscal revenue

Year	y	x1	x2	x3	x4	x5	x6	x7
2018	183359.84	380986.9	919281.1	139538	645675	102.1	156402.86	77586
2017	172592.77	366261.6	832035.9	139008	641238.4	101.6	144369.87	77640
2016	159604.97	332316.3	746395.1	138271	606465.7	102	130360.73	77603
2015	152269.23	300930.8	688858.2	137462	561999.8	101.4	124922.2	77451
2014	140370.03	271896.1	643563.1	136782	512020.7	102	119175.31	77253
2013	129209.64	242842.8	592963.2	136072	446294.1	102.6	110530.7	76977
2012	117253.52	214432.7	538580	135404	374694.7	102.6	100614.28	76704
2011	103874.43	187205.8	487940.2	134735	311485.1	105.4	89738.39	76420

2010	83101.51	158008	412119.3	134091	251683.8	103.3	73210.79	76105
2009	68518.3	133048.2	348517.7	133450	224598.8	99.3	59521.59	75828
2008	61330.35	114830.1	319244.6	132802	172828.4	105.9	54223.79	75564
2007	51321.78	93571.6	270092.3	132129	137323.9	104.8	45621.97	75321
2006	38760.2	79145.2	219438.5	131448	109998.2	101.5	34804.35	74978
2005	31649.29	68352.6	187318.9	130756	88773.6	101.8	28778.54	74647
2004	26396.47	59501	161840.2	129988	70477.4	103.9	24165.68	74264
2003	21715.25	52516.3	137422	129227	55566.6	101.2	20017.31	73736
2002	18903.64	48135.9	121717.4	128453	43499.9	99.2	17636.45	73280
2001	16386.04	43055.4	110863.1	127627	37213.5	100.7	15301.38	72797
2000	13395.23	39105.7	100280.1	126743	32917.7	100.4	12581.51	72085
1999	11444.08	35647.9	90564.4	125786	29854.7	98.6	10682.58	71394
1998	9875.95	33378.1	85195.5	124761	28406.2	99.2	9262.8	70637
1997	8651.14	31252.9	79715	123626	24941.1	102.8	8234.04	69820
1996	7407.99	28360.2	71813.6	122389	22913.5	108.3	6909.82	68950
1995	6242.2	23613.8	61339.9	121121	20019.3	117.1	6038.04	68065
1994	5218.1	18622.9	48637.5	119850	17042	124.1	5126.88	67455
1993	4348.95	14270.4	35673.2	118517	13072	114.7	4255.3	66808
1992	3483.37	10993.7	27194.5	117171	8080.1	106.4	3296.91	66152
1991	3149.48	9415.6	22005.6	115823	5594.5	103.4	2990.17	65491
1990	2937.1	8300.1	18872.9	114333	4517	103.1	2821.86	64749
1989	2664.9	8101.4	17179.7	112704	4410.4	118	2727.4	55329
1988	2357.24	7440	15180.4	111026	4753.8	118.8	2390.47	54334
1987	2199.35	5820	12174.6	109300	3791.7	107.3	2140.36	52783
1986	2122.01	4950	10376.2	107507	3120.6	106.5	2090.73	51282
1985	2004.82	4305	9098.9	105851	2543.2	109.3	2040.79	49873
1984	1642.86	3376.4	7278.5	104357	1832.9	102.7	947.35	48197
1983	1366.95	2849.4	6020.9	103008	1430.1	102	775.59	46436
1982	1212.33	2570	5373.4	101654	1230.4	102	700.02	45295
1981	1175.79	2350	4935.8	100072	961	102.5	629.89	43725
1980	1159.93	2140	4587.6	98705	910.9	107.5	571.7	42361
1979	1146.38	1800	4100.5	97542	890	101.9	537.82	41024
1978	1132.26	1558.6	3678.7	96259	879	100.7	519.28	40152
1977	874.46	1432.8	3250	94974	856	102.7	468.27	39377
1976	776.58	1339.4	2988.6	93717	812	100.3	407.96	38834
1975	815.61	1271.1	3039.5	92420	769	100.4	402.77	38168
1974	783.14	1163.6	2827.7	90859	726	100.7	360.4	37369
1973	809.67	1106.7	2756.2	89211	688	100.1	348.95	36652
1972	766.56	1023.3	2552.4	87177	666	100.2	317.02	35854
1971	744.73	929.2	2456.9	85229	610	99.9	312.56	35620
1970	662.9	858	2279.7	82992	570	100	281.2	34432

3. Descriptive Statistics

3.1. Analysis of Descriptive Statistical Outcomes

Table 3.1. Descriptive Statistical Outcomes

	y	x1	x2	x3	x4	x5	x6	x7
count	49.00	49.00	49.00	49.00	49.00	49.00	49.00	49.00
mean	34269.17	70538.44	173502.43	116365.86	112890.75	104.14	29399.32	60874.63
std	53706.64	106434.54	252347.36	17130.78	192832.09	5.58	45243.54	15616.14
min	662.90	858.00	2279.70	82992.00	570.00	98.60	281.20	34432.00
25%	1212.33	2570.00	5373.40	101654.00	1230.40	100.70	700.02	45295.00
50%	5218.10	18622.90	48637.50	119850.00	17042.00	102.10	5126.88	67455.00
75%	38760.20	79145.20	219438.50	131448.00	109998.20	105.40	34804.35	74978.00
max	183359.84	380986.90	919281.10	139538.00	645675.00	124.10	156402.86	77640.00

Firstly, computing the descriptive statistical results of data to understand the overall characteristics of the data. In general, the total amount of data are 49 records. As Table 3.1 shown to us, the mean of fiscal revenue is 34269.17, and its standard deviation is 53706.64, the statistical results of fiscal revenue shows that great difference existed throughout given years (from 1970 to 2018). Also, there are great difference between the means of explanatory variables x_1, x_2, x_4, x_6 and their own standard deviations, which shows obvious changes of statistical values in different years.

3.2. Histogram of Variables

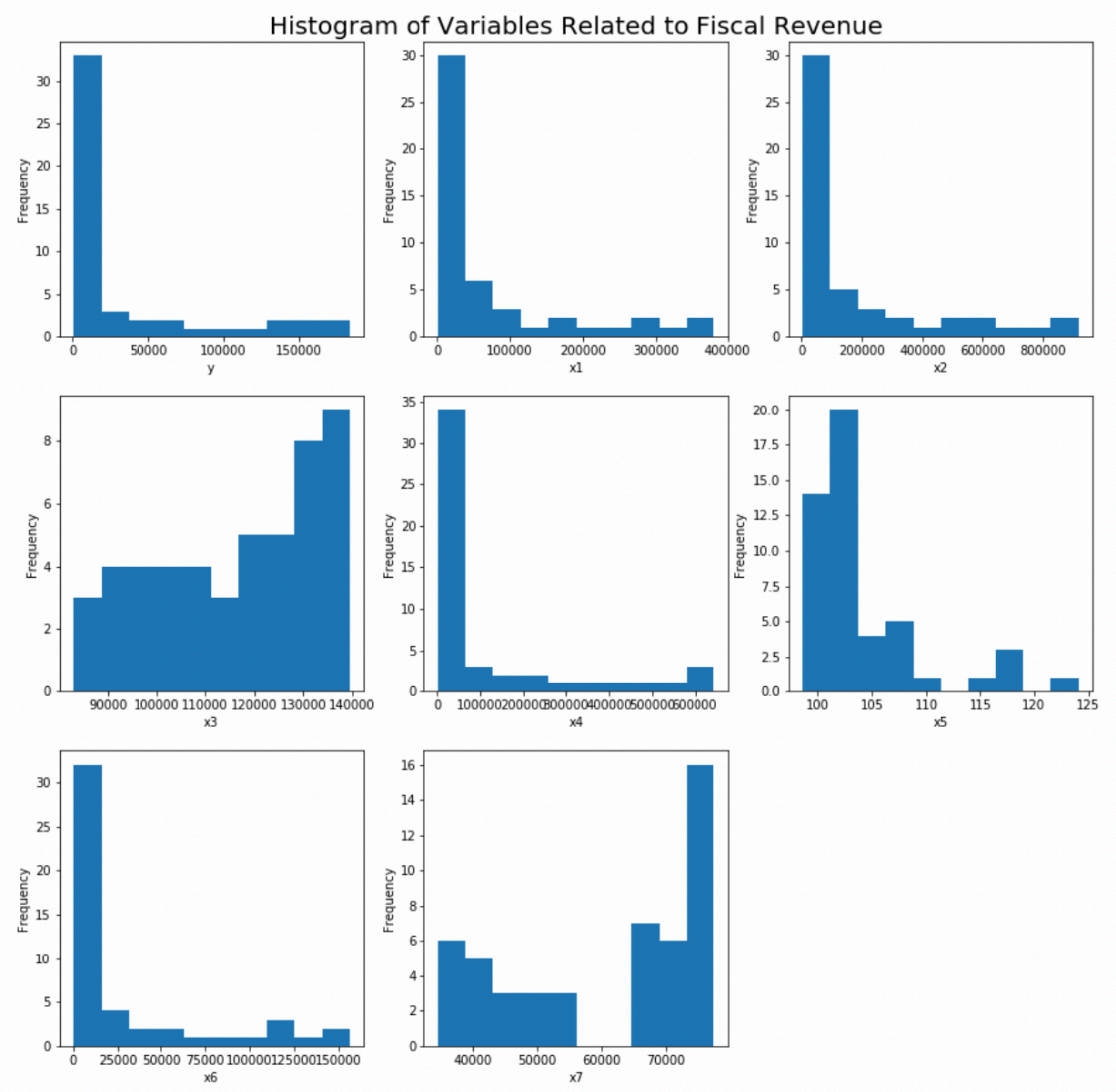

Figure 3.2 Histogram of Variables Related to Fiscal Revenue

As can be seen in the Figure 3.2, it is clear that each subplot has split the range of column into ten bins and shows their own characteristics of distribution. The target variable y shows an obvious right skewed distribution. As for explanatory variables, x_1, x_2, x_4, x_5, x_6 shows an obvious right skewed distribution; x_3, x_7 shows a left skewed distribution, but the change of x_3 is more stable.

3.3. Boxplot of Variables (Before and After Normalization)

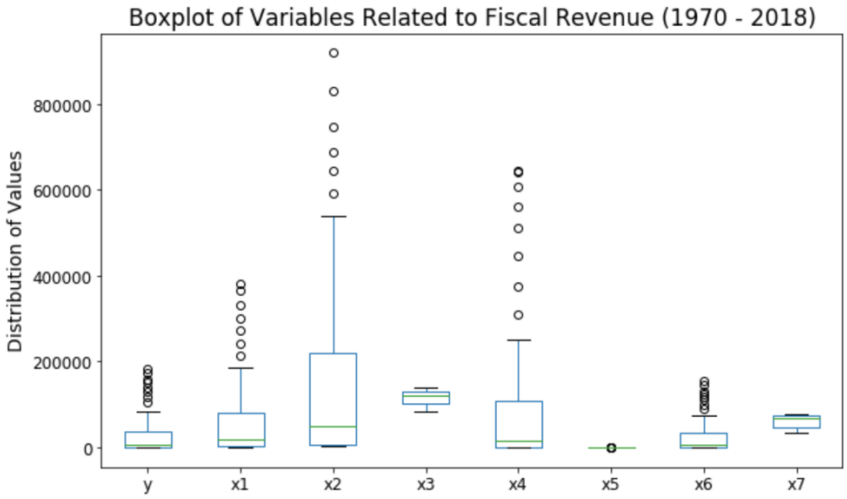

Figure 3.3.1 Boxplot of Variables Related to Fiscal Revenue (Before Normalization)

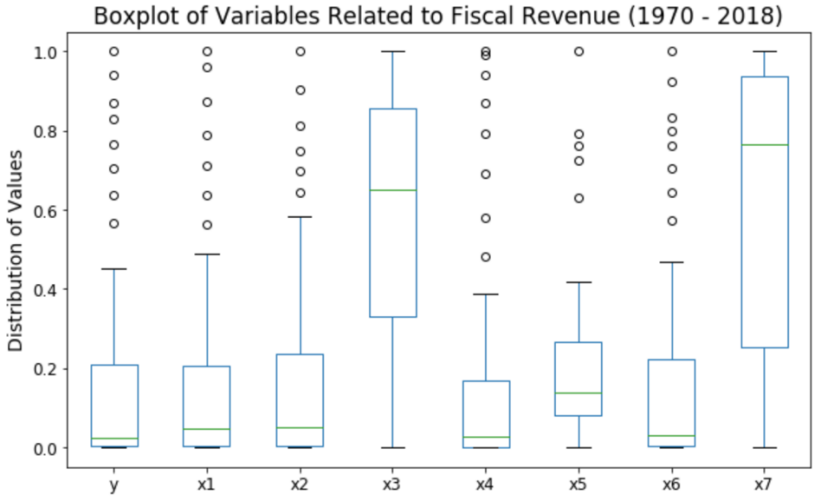

Figure 3.3.2 Boxplot of Variables Related to Fiscal Revenue (After Normalization)

Comparing the Figure 3.3.1 and Figure 3.3.2, these two boxplots show different distribution results for variables. This is because the scale of measurement can be greatly reduced and people can explore its characteristics of distribution in a more detailed and clear visualization. Therefore, the result of measurement is more accurate after normalization and we will analyze the result of boxplot after normalization. As Figure 3.3.2 shown to us, x_3, x_7 have relatively larger ranges. In addition, each variable has their own outliers that all of them are more than their maximum, thus, these variables show right skewed distribution. Obviously, this distribution characteristic illustrates that the economic development of China is in a rising trend.

4. Scatterplot and Correlation Coefficients of Variables

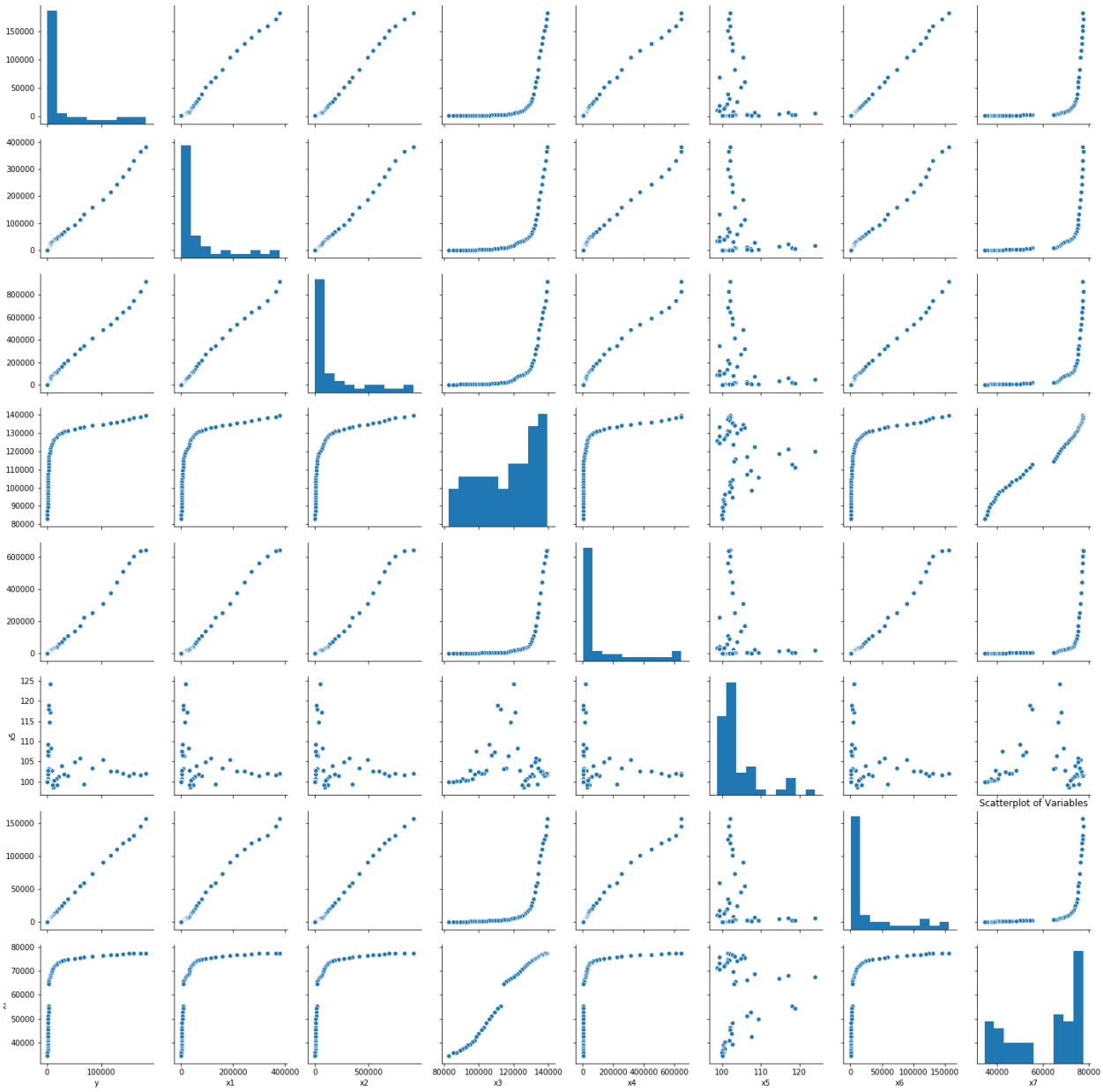

Figure 4 Scatterplot of Variables

Table 4 Correlation Coefficients of Variables

	y	x1	x2	x3	x4	x5	x6	x7
y	1.0000	0.9968	0.9971	0.7037	0.9959	-0.1828	0.9996	0.6272
x1	0.9968	1.0000	0.9976	0.7163	0.9963	-0.1793	0.9959	0.6422
x2	0.9971	0.9976	1.0000	0.7378	0.9907	-0.1809	0.9980	0.6652
x3	0.7037	0.7163	0.7378	1.0000	0.6719	0.0435	0.7158	0.9873
x4	0.9959	0.9963	0.9907	0.6719	1.0000	-0.1777	0.9931	0.5940
x5	-0.1828	-0.1793	-0.1809	0.0435	-0.1777	1.0000	-0.1810	0.0623
x6	0.9996	0.9959	0.9980	0.7158	0.9931	-0.1810	1.0000	0.6404
x7	0.6272	0.6422	0.6652	0.9873	0.5940	0.0623	0.6404	1.0000

According to the Figure 4 and Table 4, it is clear that explanatory variable x_3, x_7 have positive correlation with target variable y (with correlation coefficients 0.7037 and 0.6272 respectively); However, the explanatory variable x_5 has a weak correlation with target variable y (with correlation coefficient -0.1828); In addition, explanatory variables x_1, x_2, x_4, x_6 have a highly positive correlation with target variable y (with correlation coefficients 0.99), therefore, this situation might be assumed as multicollinearity. In statistics, Multicollinearity refers to that the explanatory variables in a linear regression model are distorted or difficult to estimate accurately due to the existence of precise or highly correlated relationships. So, we will explore appropriate method to eliminate multicollinearity and build linear regression model.

5. Multicollinearity Test and Linear Regression Model

5.1. Multicollinearity Test

In statistics, the variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity. The VIF results as the table below.

Table 5.1 Results of VIF Factor

	VIF Factor	Features
0	13.791658	Intercept
1	2700.787109	x1
2	2837.217584	x2
3	77.804872	x3
4	1067.659256	x4
5	1.205877	x5
6	1203.949313	x6
7	62.420097	x7

Based on the results of VIF factor, it is clear that the VIF factors of x_1, x_2, x_4, x_6 are larger than 100, and thus it can be seen that the multicollinearity existed among them. In order to select the best model, this study compares different R^2 of linear regression models as the table below.

Table 5.2 Comparison for Different R^2 of Linear Regression Models

No. Linear Regression Model	Independent Variables	R^2
1	x_1, x_3, x_5, x_7	0.994
2	x_2, x_3, x_5, x_7	0.997
3	x_4, x_3, x_5, x_7	0.994
4	x_6, x_3, x_5, x_7	0.999
5	x_3, x_5, x_7	0.702

Table 5.3 VIF Factor of Model 4

	VIF Factor	Features
0	13.137582	Intercept
1	3.479283	x6
2	75.303332	x3
3	1.110753	x5
4	61.726932	x7

From the table 5.2, by comparing different R^2 of linear regression models, model 4 has the highest R^2 (with 0.999). And its VIF factor can be shown in the table 5.3. Therefore, we select independent variables of model 4 to fit the linear regression model.

5.2. Build Linear Regression Model

In statistics, ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function. So, in this study, we will use OLS to obtain the regression results, as the table 5.4 below.

Table 5.4 OLS Regression Results

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.999
Model:	OLS	**Adj. R-squared:**	0.999
Method:	Least Squares	**F-statistic:**	1.846e+04
Date:	Tue, 25 Feb 2020	**Prob (F-statistic):**	2.56e-70
Time:	20:56:39	**Log-Likelihood:**	-420.76
No. Observations:	49	**AIC:**	851.5
Df Residuals:	44	**BIC:**	861.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4226.8033	6312.799	0.670	0.507	-8495.807	1.69e+04
x6	1.2072	0.008	148.201	0.000	1.191	1.224
x3	-0.0550	0.100	-0.550	0.000	-0.257	0.147
x5	23.0204	37.345	0.616	0.008	-52.244	98.285
x7	-0.0237	0.099	-0.238	0.000	-0.224	0.177

Omnibus:	28.312	**Durbin-Watson:**	0.783
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	76.443
Skew:	1.492	**Prob(JB):**	2.52e-17
Kurtosis:	8.342	**Cond. No.**	4.45e+06

Suppose the data consists of n observations $\{y_i, x_i\}_{i=1}^{n}$. Each observation i includes a scalar response y_i and a column vector x_i of values of p predictors (regressors) x_{ij} for $j = 1, 2, \ldots, p$. In a linear regression model, the response variable, y_i, is a linear function of the regressors:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Therefore, the linear regression model of fiscal revenue is:

$$\mathrm{y} = 4226.8033 - 0.055x_3 + 23.0204x_5 + 1.2072x_6 - 0.0237x_7$$

In conclusion, the main factors of affecting the fiscal revenue are *total population at the end of each year (x_3), consumer price index (x_5), tax (x_6), number of employees (x_7).*

5.3. T-Test Analysis

A t-test's statistical significance indicates whether or not the difference between two groups' averages most likely reflects a "real" difference in the population from which the groups were sampled. This analysis is appropriate whenever comparing the means of two groups, and especially appropriate as the analysis for the posttest-only two-group randomized experimental design. The final formula for the t-test is:

$$t = \frac{signal}{noise} = \frac{difference\ between\ group\ means}{variability\ of\ groups} = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{var(t)}{n_t} + \frac{var(c)}{n_c}}}$$

To test the significance, we set a risk level ($\alpha = 0.05$). This means that five times out of a hundred we would find a statistically significant difference between the means even if there was none.

From the table 5.4, OLS regression results show that the p-values of x_3, x_5, x_6, x_7 are approximately 0.000, 0.000, 0.008, 0.000 respectively, which are much lower than significance level ($\alpha = 0.05$). Therefore, we reject the null hypothesis and conclude that there is a statistically significant difference among these four independent variables with respect to fiscal revenue.

6. ANOVA and F-Test Analysis

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means. The result of ANOVA for the linear regression model of fiscal revenue can be shown in the table as below.

Table 6 ANOVA

	df	sum_sq	mean_sq	F	PR(>F)
x6	1.0	1.383286e+11	1.383286e+11	73823.626437	1.342264e-72
x3	1.0	3.945913e+07	3.945913e+07	21.058669	3.698179e-05
x5	1.0	7.603268e+05	7.603268e+05	0.405774	5.274231e-01
x7	1.0	1.064426e+05	1.064426e+05	0.056807	8.127229e-01
Residual	44.0	8.244594e+07	1.873771e+06	NaN	NaN

For one-way ANOVA, the ratio of the between-group variability to the within-group variability follows an F-distribution when the null hypothesis is true. When performing a one-way ANOVA for our study, we obtain a single F-value. Because the F-distribution assumes that the null hypothesis is true, we can place the F-value from our study in the F-distribution to determine how consistent our results are with the null hypothesis and to calculate probabilities. The probability that we want to calculate is the probability of observing an F-statistic that is at least as high as the value that our study obtained. That probability allows us to determine how common or rare our F-value is under the assumption that the null hypothesis is true. If the probability is low enough, we can conclude that our data is inconsistent with the null hypothesis. The evidence in the sample data is strong enough to reject the null hypothesis for the entire population. This probability that we're calculating is also known as the p-value.

From the table 6, it is clear that the p-values of x_3, x_5, x_6, x_7 are much lower than significance level ($\alpha = 0.05$). Therefore, we reject the null hypothesis of equal population means and conclude that there is a statistically significant difference among the population means.