复旦大学计算机科学技术学院

2015-2016 学年第一学期期末论文课程评分表

课程名称： 数据分析　　**课程代码：** COMP110041.01

开课院系： 计算机科学技术学院

学生姓名： 李仁杰　**学号：** 13307130279　**专业：** 电子信息科学与技术

论文名称： 中国 1978 年-2014 年税收数据分析

（以上由学生填写）

成绩：

论文评语（教师填写）：
任课教师签名： 日　　期：

中国1978年-2014年税收数据分析

姓名：李仁杰　　学号：13307130279

[摘 要] 财政收入是一国政府实现政府职能的基本保障，而税收是财政收入的主要组成部分。税收的增长情况和一个国家经济的发展和社会的进步有着密切的关系。我国税收主要受国民经济发展等因素的影响，尤其是国内生产总值。本文针对我国税收的影响因素建立了数学模型，并利用 Excel、MATLAB 和 R 软件对收集到的时间序列数据进行相关回归分析，排除简单多元回归模型存在的严重多重共线性等问题，建立税收影响因素更精确的模型，分析了影响税收主要因素及其影响程度，最终得到税收与国内生产总值的分段线性关系。本文建立了趋势增长模型，并预测了我国税收和国内生产总值的增长趋势。

[主题词] 线性回归 时间序列 税收 国内生产总值

[Abstract] Financial income is the basic guarantee for the government to realize the function of the government, and the tax is the main component of the revenue. The growth of tax and the development of a country's economy and the progress of the society are closely related. The tax income of our country is mainly impacted by the development of the national economy, especially the gross domestic product. This paper establishes a mathematical model for the impact factors of tax revenue. Then this paper uses Excel, MATLAB and R software to analyze the data collected from the relevant regression analysis and eliminate the problems of simple multiple regression model, and establish a more accurate model, analyze the main factors affecting tax revenue and its impact, and forecast the growth trend of China's tax revenue and GDP.

1.导言和数据处理

据统计1978～2014年我国税收（TAX）的规模随着经济的不断增长而增长，由1978 年的519.28亿元到2014年的119158.05亿元,扩大了近200倍。据《财经网》报道[1]，早在2006-2010的“十一五”期间，财政收入年度增速已经数倍于同期GDP。同时，“中国税负是否过高”引发了全民关注。

为了研究影响中国税收增长的主要原因，预测中国财政收入未来的增长趋势，需要建立合适的计量经济学模型进行数据分析。影响中国财政收入增长的因素很多，但据分析，主要的因素为经济发展水平。经济发展水平的影响是基础性的，经济发展水平与财政收入是根与叶、源与流的关系。在本篇分析报告中，经济发展水平用国内生产总值(GDP),进出口贸易总额(IE),社会消费品零售总额(RS)，城乡居民消费额（COM)，全社会固定资产投资总额（INV)，城乡居民储蓄存款年底余额（DEP）来体现。同时，税收（TAX）也取决于国家的财政支出（EXP)，故同样加入量化。综上所诉，我们可以从以上几个方面，分析各种因素对中国税收增长的具体影响。显而易见，国内生产总值（GDP）在经济发展水平中较其他的数据更为显著。

分析用到的数据如附录中的“data.csv”文件所示，较为原始的数据在“税收.xls”中给出，其中除城乡居民消费额(COM)之外的数据来自国家数据网[2]，

城乡居民消费额（COM）的数据来自Wind资讯[3]，由于Wind资讯中2013年和2014年的数据未经过校准，故不进行采用，设为缺省值。（单位均为百亿元）

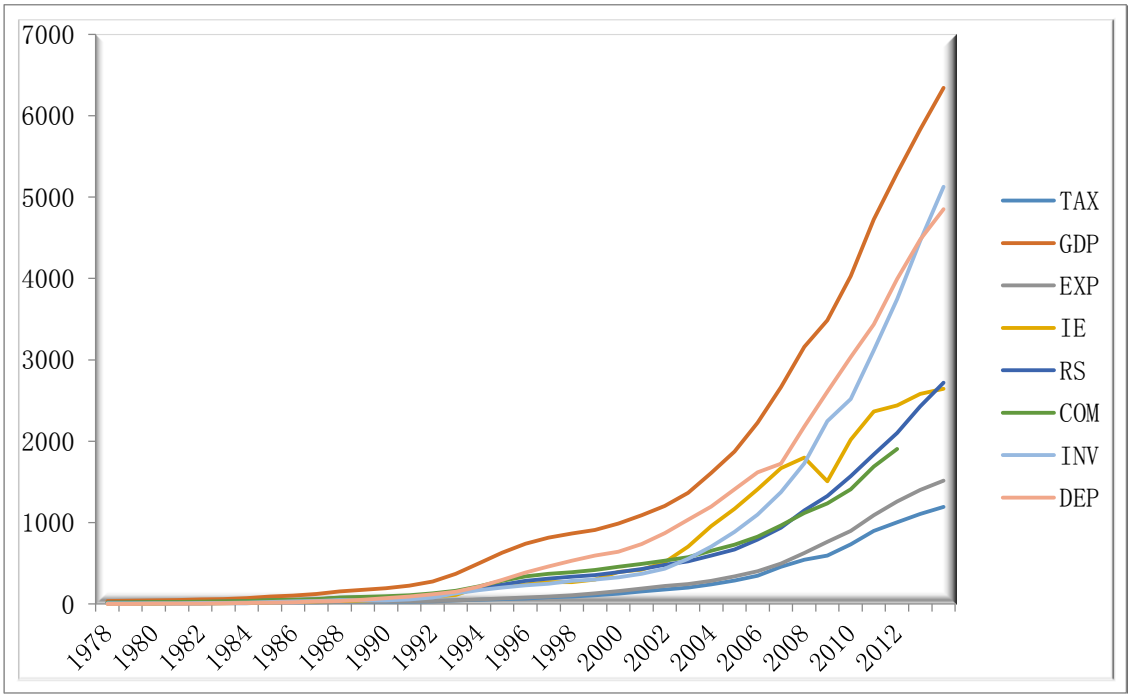

图1. 时间序列图

用Excel对这部分时间序列数据绘图如上图，我们可以发现，我们的8组数据除进出口贸易总额（IE）外均为递增数据，其中由于2008年的金融危机，进出口贸易总额(IE)受到不可避免的负面影响。为了观察各个具体因素对税收的影响，本篇报告将采用线性回归的方法进行分析。另外，为了预测税收（TAX）和国内生产总值（GDP）在2015年的数值，我们将采用时间序列模型。

2. 分析方法

在这篇分析报告中，我们主要采用了两种分析方法，一种是线性回归（包括一元线性回归和多元线性回归），另一种是趋势移动平均法的时间序列模型。以下是有关这两种方法的原理和公式证明。

2.1 一元线性回归

一元线性回归的模型为

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (1)$$

式中，β_0，β_1为回归系数，ε是随机误差项，总是假设$\varepsilon \sim N(0, \sigma^2)$，则随机变量$y \sim N(\beta_0 + \beta_1 x, \sigma^2)$。

若对y和x分别进行了n次独立观测，得到以下n对观测值

$$(y_i, x_i), \ i = 1,2,\cdots,n \quad (2)$$

这n对观测值之间的关系符合模型

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,2,\cdots,n \quad (3)$$

这里，x_i是自变量在第i次观测时的取值，它是一个非随机变量，并且没有测量误差。对应于x_i，y_i是一个随机变量，它的随机性是由ε_i造成的。$\varepsilon \sim N(0, \sigma^2)$，

对于不同的观测，当$i \neq j$时，ε_i与ε_j是相互独立的。

我们用最小二乘法估计β_0, β_1的值，即取β_0, β_1的一组估计值$\hat{\beta}_0$，$\hat{\beta}_1$，使y_i与$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$的误差平方和达到最小。若记

$$\mathrm{Q}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

则

$$\mathrm{Q}(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0,\ \beta_1} \mathrm{Q}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

显然$\mathrm{Q}(\beta_0, \beta_1) \geq 0$，且关于$\beta_0, \beta_1$可微，则由多元函数存在极值的必要条件得

$$\frac{\partial \mathrm{Q}}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \mathrm{Q}}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

整理后，得到下面的方程组

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{4}$$

此方程组称为正规方程组，求解可以得到

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases} \tag{5}$$

称$\hat{\beta}_0$，$\hat{\beta}_1$为β_0, β_1的最小二乘估计，其中，$\bar{x}$，$\bar{y}$分别是x_i与y_i的样本均值，即

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ , \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

2.2 多元线性回归

多元线性回归分析的模型为

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \tag{6}$$

式中$\beta_0, \beta_1, \cdots, \beta_m, \sigma^2$都是与$x_1, x_2, \cdots, x_m$无关的未知参数，其中$\beta_0, \beta_1, \cdots, \beta_m$称为回归系数。

现得到n个独立观测数据$(y_i, x_{i1}, \cdots, x_{im})$，$i = 1,2, \cdots, n, n > m$，由（6）得

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2), i = 1,2, \cdots, n \end{cases} \tag{7}$$

记

$$\mathbf{X}=\begin{bmatrix}1 & x_{11} & \cdots & x_{1m}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{nm}\end{bmatrix},\ \mathbf{Y}=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix} \tag{8}$$

$$\boldsymbol{\varepsilon}=[\varepsilon_1 \quad \cdots \quad \varepsilon_n]^T, \boldsymbol{\beta}=[\beta_0 \quad \beta_1 \quad \cdots \quad \beta_m]^T$$

（6）表为

$$\begin{cases}\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}\sim N(0,\sigma^2\mathbf{E_n})\end{cases} \tag{9}$$

其中$\mathbf{E_n}$为n 阶单位矩阵。

模型（6）中的参数$\beta_0,\beta_1,\cdots,\beta_m$仍用最小二乘法估计，我们采用与一元线性回归类似的参数估计方式，最终得到：当矩阵$\mathbf{X}$列满秩时，$\mathbf{X^TX}$为可逆方阵，我们有如下估计：

$$\widehat{\boldsymbol{\beta}}=(\mathbf{X^TX})^{-1}\mathbf{X^TY} \tag{10}$$

将$\widehat{\boldsymbol{\beta}}$代回原模型得到y的估计值

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_mx_m \tag{11}$$

而这组数据的拟合值为$\widehat{\mathbf{Y}}=\mathbf{X}\widehat{\boldsymbol{\beta}}$，拟合误差$\mathbf{e}=\mathbf{Y}-\widehat{\mathbf{Y}}$称为残差，可作为随机误差$\boldsymbol{\varepsilon}$的估计，而

$$\mathrm{Q}=\sum_{i=1}^{n}{e_i}^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

为残差平方和（或剩余平方和），即$\mathrm{Q}(\hat{\beta})$。

2.3 时间序列

在时间序列没有明显的趋势变动时，我们可以用简单移动平均法和加权移动平均法进行数据的预测。但当时间序列出现直线增加或减少的变动趋势时，例如本篇分析报告中的GDP数据、税收数据以及其他数据均为递增数据，用简单移动平均法和加权移动平均法来预测就会出现滞后偏差。因此，需要进行修正，修正的方法是作二次移动平均，利用移动平均滞后偏差的规律来建立直线趋势的预测模型。这就是趋势移动平均法。

设观测序列为$y_1,\cdots,y_T$，取移动平均的项数$\mathrm{N}<\mathrm{T}$。一次移动的平均数为$M_t^{(1)}=\frac{1}{N}(y_t+y_{t-1}+\cdots+y_{t-N+1})$。在一次移动平均的基础上再进行一次移动平均就是二次移动平均，其计算公式为

$$M_t^{(2)}=\frac{1}{N}\left(M_t^{(1)}+\cdots+M_{t-N+1}^{(1)}\right)=M_{t-1}^{(2)}+\frac{1}{N}(M_t^{(1)}-M_{t-N}^{(1)}) \tag{12}$$

下面讨论如何利用移动平均的滞后偏差建立直线趋势预测模型。

设时间序列$\{y_t\}$从某时期开始具有直线趋势，且认为未来时期也按此直线趋势变化，则可设此直线趋势预测模型为

$$\hat{y}_{t+T}=a_t+b_tT, T=1,2,\cdots \tag{13}$$

其中t为当前时期数；T为由t至预测期的时期数；a_t为截距；b_t为斜率。两者又称为平滑系数。

现在，我们根据移动平均值来确定平滑系数。由模型（13）可知

$$\begin{cases} y_t = a_t \\ y_{t-1} = y_t - b_t \\ y_{t-2} = y_t - 2b_t \\ \cdots \\ y_{t-N+1} = y_t - (N-1)b_t \end{cases}$$

所以

$$M_t^{(1)} = \frac{y_t + y_{t-1} + \cdots + y_{t-N+1}}{N} = \frac{y_t + (y_t - b_t) + \cdots + [y_t - (N-1)b_t]}{N}$$

因此

$$y_t - M_t^{(1)} = \frac{N-1}{2} b_t \qquad (14)$$

由式（13），类似式（14）的推导，可得$y_{t-1} - M_{t-1}^{(1)} = \frac{N-1}{2} b_t$

所以$y_t - y_{t-1} = M_t^{(1)} - M_{t-1}^{(1)} = b_t$

类似式（14）的推导，可得$M_t^{(1)} - M_t^{(2)} = \frac{N-1}{2} b_t$

于是，可得平滑系数的计算公式为

$$\begin{cases} a_t = 2M_t^{(1)} - M_t^{(2)} \\ b_t = \frac{2}{N-1}(M_t^{(1)} - M_t^{(2)}) \end{cases} \qquad (15)$$

趋势移动平均法对于同时存在直线趋势与周期波动的序列，是一种既能反映趋势变化，又可以有效地分离出来周期变动的方法。

3. 统计建模与分析

3.1 GDP和税收的一元线性回归

首先，单独把国内生产总值拿出来与税收进行线性回归分析。对于税收（TAX）随国内生产总值（GDP）变化，可建立数学模型进行回归分析。在此假设拟建立如下的一元回归模型：

$$\mathrm{TAX} = \beta_0 + \beta_1 GDP + \varepsilon \qquad (16)$$

其中β_0和β_1是参数，称为回归系数，由模型可知被解释变量税收（TAX）除了受解释变量国内生产总值（GDP）的系统性影响外，还受其他诸多因素的随机性影响，ε即为这些影响因素的代表。

我们采用MATLAB对数据进行回归分析（代码见附录中的“analysis.m”文件），根据程序计算结果，税收（TAX）与国内生产总值（GDP）可建立如下关系式：

$$\mathrm{TAX} = -29.2610 + 0.1888 * \mathrm{GDP}$$

其中$\hat{\beta}_0 = 0.1888$，$\hat{\beta}_1 = -29.2610$，$\hat{\beta}_0$的置信区间是[0.1832,0.1943]，$\hat{\beta}_1$的置信区间是$[-41.8824, -16.6396]$，$R^2 = 0.9927$。绘制出的拟合曲线如图2。

从回归估计的结果看，模型拟合较好：可决系数$R^2 = 0.9927$,斜率β_1的置信区间也较小，说明变量GDP是显著的，拟合情况较好。斜率项0.1888表明，国内生产总值（GDP）每增加1亿元时，税收（TAX）平均增加0.1888亿元。我们用MATLAB

的rcoplot(r,rint)命令绘制出图3的残差分布，发现拟合残差总体较小（绘制图形上未出现明显偏离的红色）。通过对模型的检验可知税收与国内生产总值的相关性较好，被解释变量税收（TAX）受解释变量国内生产总值（GDP）的影响大，受其他因素的影响较小。

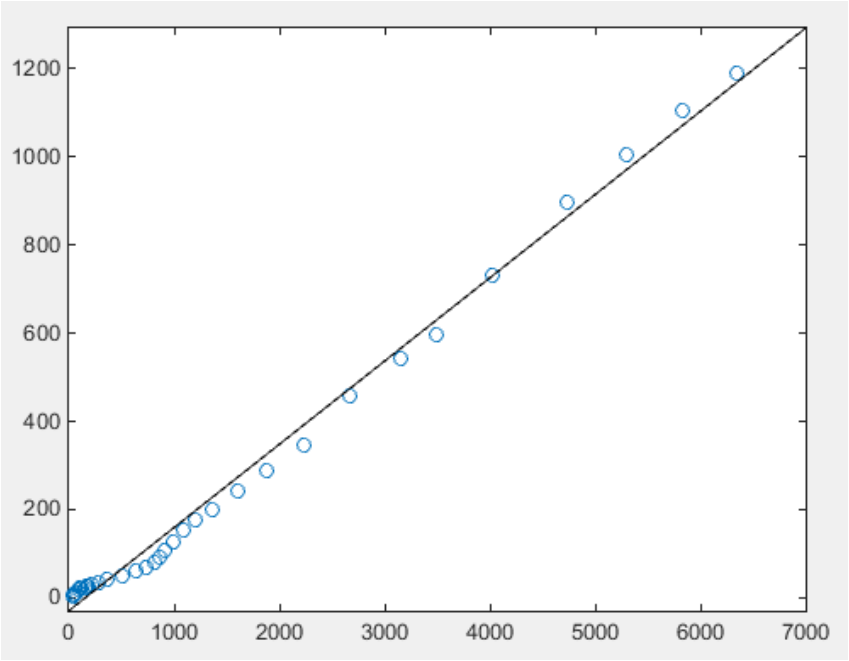

图2.GDP~TAX的一元线性回归

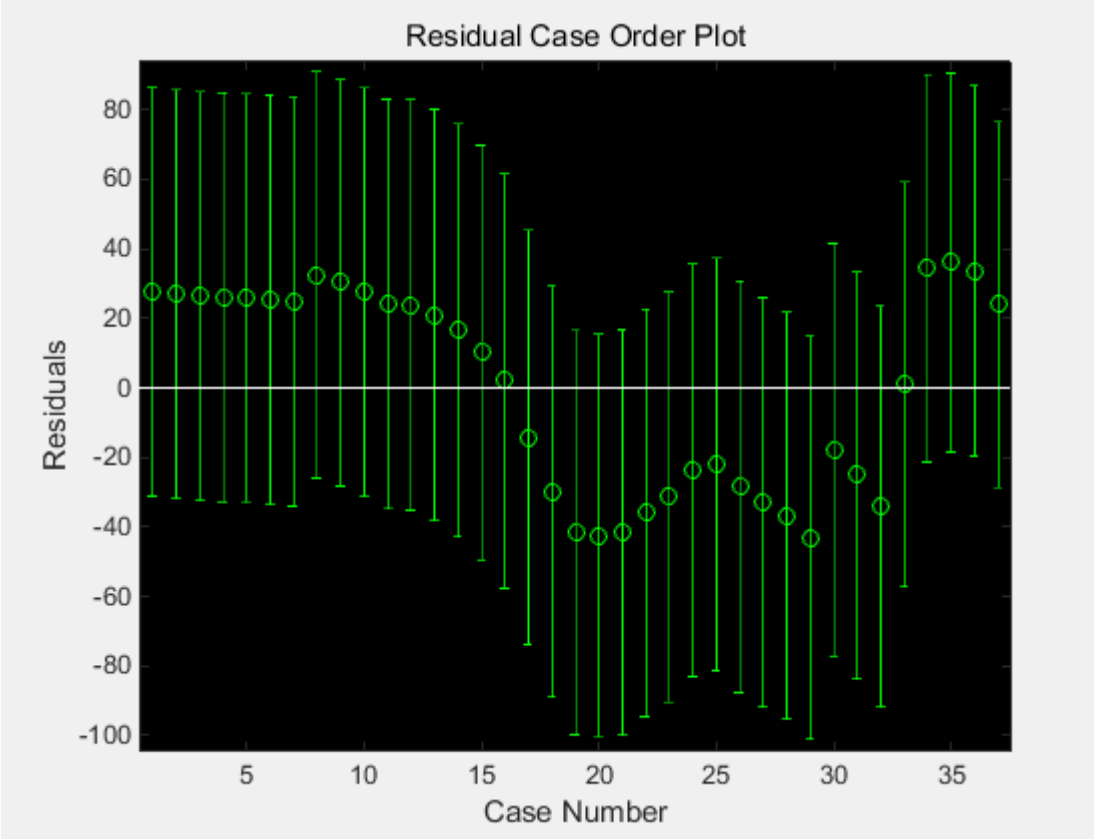

图3.GDP~TAX线性回归的残差分布

3.2 含虚拟变量的回归模型

虽然在3.1中的一元线性回归模型的拟合效果比较好，但是如果用R语言绘制出（3.2小节中所有的R语言的代码均在附录的“figure.R”文件中）国内生产总值（GDP）和税收（TAX）在时间（T）之下的变化（如图4的上面两张图），就可以明显的看到，自从1993年开始，国内生产总值（GDP）和税收（TAX）都有大幅度提升，1993年以前的国内生产总值（GDP）对税收的影响明显不如1993年以后的影响。同时我们再观察图3中的残差分布，可以发现残差分布在前半段出现了先上升后下降的趋势，而不是随机的波动，这说明一段直线拟合的功能有待商榷。而这经济现象背后是有历史原因的，1992年春，邓小平发表南巡讲话，回答了“姓‘社’还是姓‘资’”的问题，同年10月，中国共产党召开第十四次全国代表大会，正式确立了建立社会主义市场经济体制的改革目标。[4]自从中央的决议下来，中国的经济便走上了快车道。

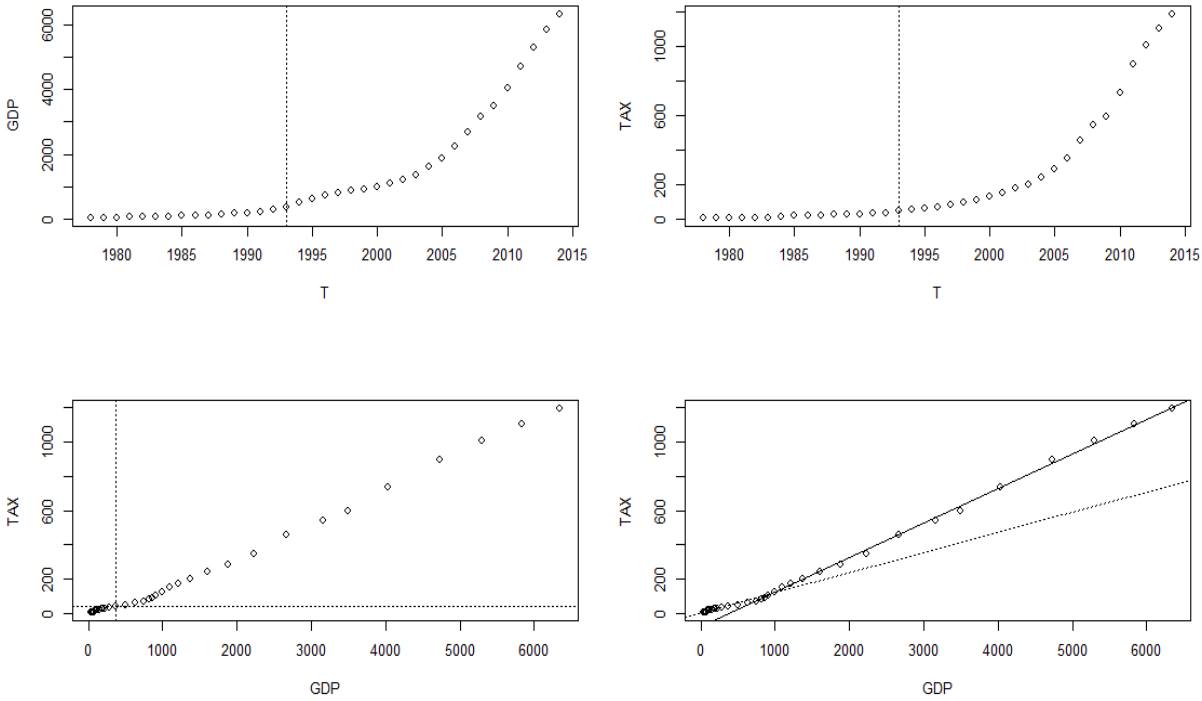

图4. 含虚拟变量的GDP~TAX回归图

由于经济数据的变化受到些许政治因素的影响，这实际上成为了一种有偏估计。有偏估计是一种把真正的因果与外部因素的影响混合在一起的估计。实证经济学在估计两个事件之间的因果关系时，要有说服力，就应消除偏差。[5]故我们引入一个虚拟变量来来检验两个不同时间段的模型是否存在显著差异。我们依照时间变量T（1993年为转折期）可以设如下虚拟变量：

$$\mathrm{D}=\begin{cases}0 & T\le 1993\\ 1 & T>1993\end{cases} \tag{17}$$

我们依然采用R语言进行分析，但是将原来的$\mathrm{TAX}\sim\mathrm{GDP}$的分析改为$\mathrm{TAX}\sim\mathrm{GDP}+\mathrm{GDP}*\mathrm{D}$的分析。通过调用R语言中的summary函数，我们得到了对税收（TAX）和国内生产总值（GDP）的分段拟合函数：

$$\mathrm{TAX}=\begin{cases}3.1137+0.1179*GDP, & T\le 1993, D=0\\ -76.39909+0.20106*GDP, & T>1993, D=1\end{cases} \tag{18}$$

同时，用R语言绘制出分段拟合的直线图（如图4右下角的图）。我们可以清晰的观察到，进行分段拟合之后，两段直线的斜率相差较大。对应于宏观的经济

中，1993年之后的经济增速也大约为之前的两倍。

3.3 对于宏观经济数据的多元线性回归

早在第一部分的图1我们便可以发现，部分宏观数据有指数增长的表象。首先我们应当对数据进行对数变换。于是我们用R语言把原数据变为时间序列数据，并调用ts.plot(log(ts(mydata,start=1978)))指令，绘制出图5。

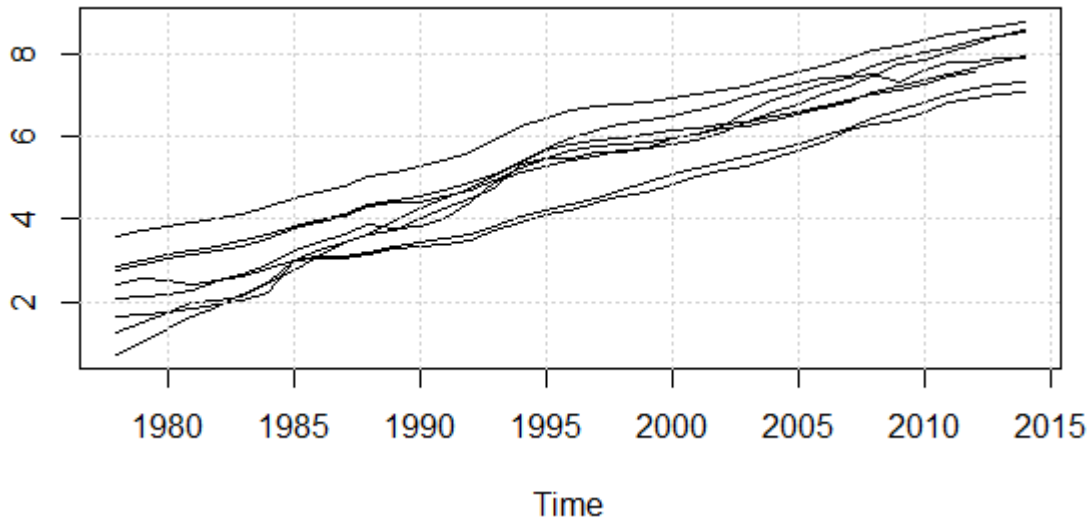

图5. 对数时间序列图

从图5可以看出，取对数后，序列基本上呈线性趋势。于是，接下来的分析我们采用是否取对数的方法进行分析。经过多次尝试，计算出多元线性回归模型的参数，最终我们发现log(TAX) ~ log(GDP) + log(EXP) + log(IE) + log(RS) + log(COM) + INV + DEP的误差最小，为最理想的模型。只需要在R语言中调用summary(lm(log(TAX)~log(GDP)+log(EXP)+log(IE)+log(RS)+log(COM)+INV+DEP,data=mydata))指令，就可以得到有关参数的估计如图6。

```
Call:
lm(formula = log(TAX) ~ log(GDP) + log(EXP) + log(IE) + log(RS) +
    log(COM) + INV + DEP, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.133331 -0.039598 -0.001819  0.032349  0.214276

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6968693  0.4363234   1.597 0.121874
log(GDP)    -1.4168695  0.3999162  -3.543 0.001463 **
log(EXP)     0.8125051  0.1659106   4.897 4.01e-05 ***
log(IE)      0.4873666  0.0958293   5.086 2.42e-05 ***
log(RS)      2.1741757  0.4864623   4.469 0.000127 ***
log(COM)    -0.9362827  0.5829792  -1.606 0.119901
INV          0.0006070  0.0002432   2.496 0.018986 *
DEP         -0.0007703  0.0002975  -2.589 0.015311 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07735 on 27 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9977
F-statistic:  2117 on 7 and 27 DF,  p-value: < 2.2e-16
```

图6. 多元线性回归参数

其中$R^2 = 0.9982, p - value < 2.2e - 16$，比3.1中的一元线性回归模型$R^2 = 0.9927$效果更好。

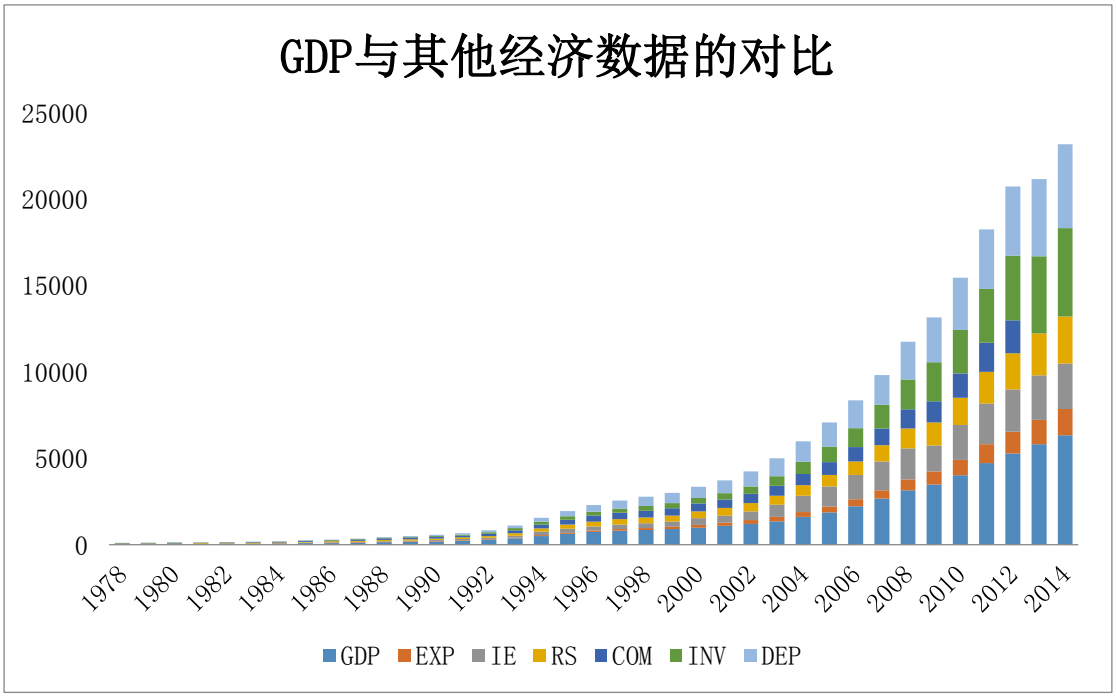

图7. GDP与其他经济数据的对比

另外我们注意到，国内生产总值（GDP）在一定程度上和财政支出（EXP），进出口贸易总额（IE），社会消费品零售总额（RS），城乡居民消费额（COM），全社会固定资产投资总额（INV），城乡居民储蓄存款年底余额（DEP）存在多重共线性。于是我们对国内生产总值和其他六个变量进行回归分析，利用R语言中的R=(summary(lm(GDP~EXP+IE+RS+COM+INV+DEP))$r.sq)^2指令，得到$R^2 = 0.9994$，这个值更为接近1。故国内生产总值（GDP）与其他变量之间存在非常严重的复共线性。这与我们的经验相符，国内生产总值（GDP）决定了其他变量的值，然后财政支出（EXP）相当于国家进行投资，拉动了内需，进出口需要支付关税，消费品零售和社会固定资产投资也同时贡献了一定的税收，在一定的时间，居民的存款的利息也需要缴税——这一切都直接增加了税收。所以说，国内生产总值（GDP）的增长是主因，其他具体的变化是直接诱因，都在一定程度上导致了税收（TAX）的增加。

3.4 趋势移动平均法的时间序列模型

通过上面对过往数据进行的分析，我们发现，影响税收（TAX）的主要因素为国内生产总值（GDP），而且在近期几年，这两项数据的的变化有一种增长的固定趋势。于是我们接下来进行对2015年的数据进行预测（2016年及以后的数据由于时间过长，可能预测会有较大误差，故不进行考虑）。

年份	t	TAX	一次平均移动	二次平均移动
1978	1	5.1928		
1979	2	5.3782		
1980	3	5.717		
1981	4	6.2989		
1982	5	7.0002		
1983	6	7.7559	6.2238	

1984	7	9.4735	6.9373	
1985	8	20.4079	9.4422	
1986	9	20.9073	11.974	
1987	10	21.4036	14.4914	
1988	11	23.9047	17.3088	11.0629
1989	12	27.274	20.5618	13.4526
1990	13	28.2186	23.686	16.244
1991	14	29.9017	25.2683	18.8817
1992	15	32.9691	27.2786	21.4325
1993	16	42.553	30.8035	24.1512
1994	17	51.2688	35.3642	27.1604
1995	18	60.3804	40.8819	30.5471
1996	19	69.0982	47.6952	34.5486
1997	20	82.3404	56.435	39.7431
1998	21	92.628	66.3781	46.2597
1999	22	106.8258	77.0903	53.9741
2000	23	125.8151	89.5147	62.9992
2001	24	153.0138	104.9535	73.6778
2002	25	176.3645	122.8313	86.2005
2003	26	200.1731	142.47	100.5397
2004	27	241.6568	167.3082	117.3613
2005	28	287.7854	197.4681	137.4243
2006	29	348.0435	234.5062	161.5896
2007	30	456.2197	285.0405	191.6041
2008	31	542.2379	346.0194	228.8021
2009	32	595.2159	411.8599	273.7004
2010	33	732.1079	493.6017	328.0826
2011	34	897.3839	595.2015	394.3715
2012	35	1006.143	704.8847	472.7679
2013	36	1105.307	813.0659	560.7722
2014	37	1191.581	921.2897	656.6505

表1.　我国税收及一、二次移动平均值计算表

考虑到中国的政治周期为5年(一届国家主席、国务院总理的任期为5年)，经济周期也将与此类似，为了保证预测的准确性，我们选取$\mathrm{N}=6$，分别计算一次和二次移动平均值如表1。

$$M_{37}^{(1)}=921.2897, M_{37}^{(2)}=656.6505$$

再由公式（15），得

$$a_{37}=2M_{37}^{(1)}-M_{37}^{(2)}=1185.9$$

$$b_{37}=\frac{2}{6-1}\left(M_{37}^{(1)}-M_{37}^{(2)}\right)=105.8556$$

于是，得$\mathrm{t}=37$时直线趋势预测模型为

$$\hat{y}_{37+T}=1185.9+105.8556\mathrm{T}$$

预测2015年和2016年的税收（TAX）为

$$\hat{y}_{2015} = \hat{y}_{38} = \hat{y}_{37+1} = 1291.8$$

同理我们可以得到国内生产总值（GDP）在2015年的估计为6736（百亿元），其中间的计算数据不再列出，而是可视化如图8。这部分计算的 MATLAB 程序代码见附录中的“analysis.m”文件后半部分。当然只要改变相应的程序段，我们也可以通过同样的方法求出其他宏观数据的预测值。

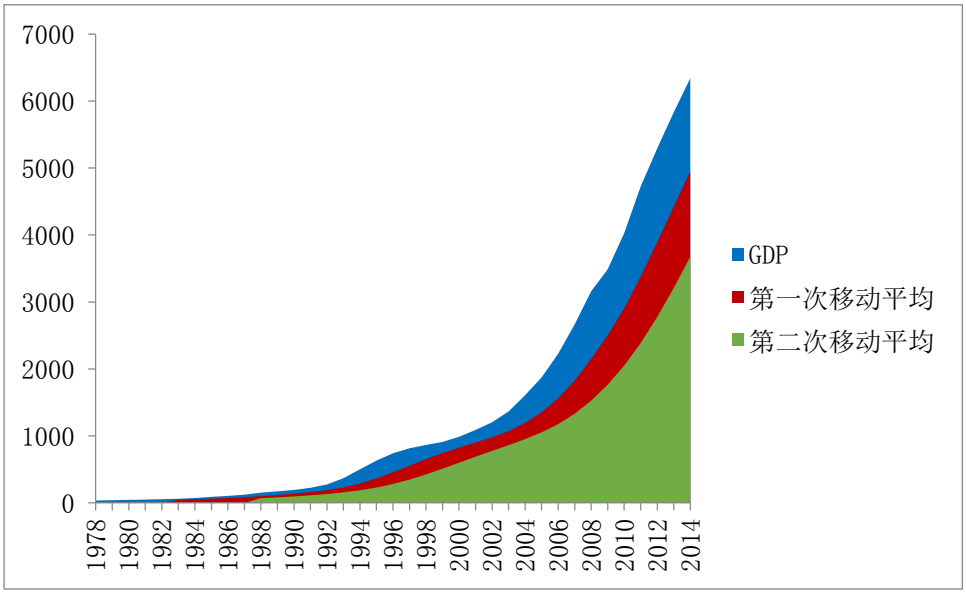

图8.　我国国内生产总值及一、二次移动平均值

我们查阅国家数据网[2]，可以得到从2013至今的分季度国内生产总值（数据见附录文件“税收.xls”中第三个sheet“GDP分季度”），并用Excel作图如图9，我们可以看到国内生产总值（GDP）的分季度值有这样的特点：在一年以内，一个季度的国内生产总值（GDP）均大于前一季度，但是每一年的第一季度的国内生产总值（GDP）远低于上一年最后一季度的国内生产总值（GDP）。

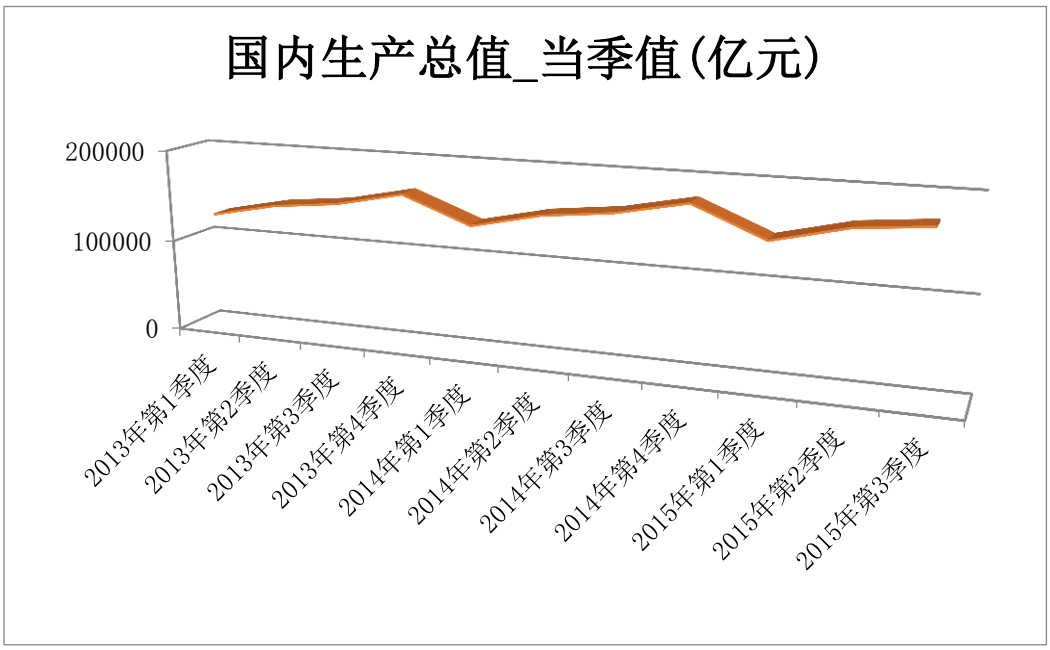

图9. GDP分季度

我们单独拿出来2015年前三个季度的国内生产总值(GDP)做累加为4877.735（百亿元),如果保持同期的增长速度,按照2014年第四季度国内生产总值(GDP)是2013年第四季度的7.55%的增速，2015年第四季度的国内生产总值应为1922.296（百亿元），于是今年的国内生产总值应为6800（百亿元）。与前面我们用趋势移动平均法做出的估计6736(百亿元)较为相近,误差处于允许范围内。我们用趋势移动平均法做出的估计值与2014年实际国内生产总值(GDP)做运算,发现增速大致为6.2%，比前些年的增速略小，这说明中国在经济转型期受到了一定的阻碍，稍有经济下行压力。

4.结论与展望

经上述论证，税收（TAX）主要与国内生产总值（GDP）成正相关关系。而且，从经济意义上面分析，虽然税收一般与其他宏观经济发展数据的关系不是很大，但还是有一定关系的，且经过多元线性回归进行了分析。

随着国家机器的不断强化、国家之间战争的发生、生产社会化所带来的国家职能扩大以及其他原因,使各国政府开支迅速增加,于是开辟多种财政收入来源。但是，在历代各国的财政收入中，税收却一直保持着它的主导地位不变。税收来源的充沛与否，至今仍然是衡量各国财政基础是否稳固的一个重要标志。由于税收与社会再生产中生产、交换、分配、消费各环节息息相关，联系密切，并且直接调节着各种经济成分的收入，影响到各个经济主体的切身利益，广泛地渗透到社会经济生活的各个方面,因此,它还是国家财政反作用于经济的一个重要杠杆。[5]财政的经济杠杆作用有相当一部分是通过税收体现的。不可否认，税收是以法律的形式存在，并且有其自身运行和发展变化的规律，不是其他财政收入形式所可比拟的。

税收作为社会生产力发展到一定阶段的产物，必然随着社会的发展而扩大。税收是国家参与一部分社会产品或国民收入分配与再分配所进行的经济活动,因此税收从一定程度上决定了国家的健康稳定发展,我国目前正处于经济体制转型期，市场机制还不完善，宏观方面，需要政府进行积极的宏观调控，实现产业结构调整，以及财政支出政策的改进。另外，我国应实行结构性减税，结合推进税制改革，用减税、退税或抵免的方式减轻税收负担，促进企业投资和居民消费，实行积极财政政策，促进国民经济稳健发展，从而对税收形成良性的影响。

参考文献

[1] http://www.caijing.com.cn/2010-12-22/110599327.html,《财经网》，中国今年财政收入超“8 万亿”成定局。

[2] http://data.stats.gov.cn/index.htm，国家数据网。

[3] http://www.wind.com.cn/，Wind 资讯--中国领先的金融数据和分析工具服务商。

[4] 张静如:《中国共产党通史（插图本)》,第三卷（下),广东人民出版社 2002 年版。

[5] 哈维·S·罗森，特德·盖亚:《财政学》，郭庆旺，赵志耘 译，北京：人

民大学出版社，2009 年第八版。

致谢

首先，感谢计算机科学与技术学院的徐志平老师能给我们这些本科生开一门内容丰富的选修课。这门课让我掌握了利用 Excel 软件和 MATLAB 软件的一些有关于数据处理、数据分析、数据可视化的技能，同时也拓宽了我的视野。在这次的课程报告中，巧妙地把我的第二专业经济学结合了进来，是我获得了新的技能，这也算是学科交叉的魅力吧。

再者，在这门课的进行过程中我结识了一些同学，我与他们在课堂上谈笑风生，和他们交流的过程也是我们思维碰撞的过程。我由此学到了很多新知识，感谢他们能耐心和我交流。

最后，在论文的写作过程我中参阅了大量的国内外数据、资料、文献。在此，谨向这些文章的作者和数据的提供者表示感谢，他们完成的工作是我课程报告写作的基础。

附录

data.csv

（单位：百亿元）

	TAX	GDP	EXP	IE	RS	COM	INV	DEP
1978	5.1928	36.056	11.2209	3.55	15.586	17.591	8.008	2.106
1979	5.3782	40.926	12.8179	4.546	18	20.115	8.565	2.81
1980	5.717	45.929	12.2883	5.7	21.4	23.312	9.109	3.958
1981	6.2989	50.088	11.3841	7.353	23.5	26.279	9.61	5.237
1982	7.0002	55.9	12.2998	7.713	25.7	29.029	12.304	6.754
1983	7.7559	62.162	14.0952	8.601	28.494	32.311	14.301	8.925
1984	9.4735	73.627	17.0102	12.01	33.764	37.42	18.329	12.147
1985	20.4079	90.767	20.0425	20.667	43.05	46.874	25.432	16.226
1986	20.9073	105.085	22.0491	25.804	49.5	53.021	31.206	22.385
1987	21.4036	122.774	22.6218	30.842	58.2	61.261	37.917	30.814
1988	23.9047	153.886	24.9121	38.218	74.4	78.681	47.538	38.222
1989	27.274	173.113	28.2378	41.56	81.014	88.126	44.104	51.964
1990	28.2186	193.478	30.8359	55.601	83.001	94.509	45.17	71.196
1991	29.9017	225.774	33.8662	72.258	94.156	107.306	55.945	92.449
1992	32.9691	275.652	37.422	91.196	109.937	130.001	80.801	117.573
1993	42.553	369.381	46.423	112.71	142.704	164.121	130.723	152.035
1994	51.2688	502.174	57.9262	203.819	186.229	218.442	170.421	215.188
1995	60.3804	632.169	68.2372	234.999	236.138	283.697	200.193	296.623
1996	69.0982	741.636	79.3755	241.338	283.602	339.559	229.135	385.208
1997	82.3404	816.585	92.3356	269.672	312.529	369.215	249.411	462.798

1998	92.628	865.316	107.9818	268.497	333.781	392.293	284.062	534.075
1999	106.8258	911.25	131.8767	298.962	356.479	419.204	298.547	596.218
2000	125.8151	987.49	158.865	392.732	391.057	458.546	329.177	643.324
2001	153.0138	1090.28	189.0258	421.836	430.554	494.359	372.135	737.624
2002	176.3645	1204.756	220.5315	513.782	481.359	530.566	434.999	869.1065
2003	200.1731	1366.134	246.4995	704.835	525.163	576.498	555.666	1036.173
2004	241.6568	1609.566	284.8689	955.391	595.01	652.185	704.774	1195.554
2005	287.7854	1874.234	339.3028	1169.218	671.766	729.587	887.736	1410.51
2006	348.0435	2227.125	404.2273	1409.74	791.452	825.7545	1099.982	1615.873
2007	456.2197	2665.992	497.8135	1668.637	935.716	963.325	1373.239	1725.342
2008	542.2379	3159.746	625.9266	1799.215	1148.301	1116.704	1728.284	2178.854
2009	595.2159	3487.751	762.9993	1506.481	1326.784	1235.846	2245.988	2607.717
2010	732.1079	4028.165	898.7416	2017.222	1569.984	1407.587	2516.838	3033.025
2011	897.3839	4726.192	1092.478	2364.02	1839.186	1689.566	3114.851	3436.359
2012	1006.143	5293.992	1259.53	2441.602	2103.07	1905.846	3746.947	3995.51
2013	1105.307	5831.967	1402.121	2581.689	2428.428	NA	4462.941	4476.016
2014	1191.581	6340.434	1516.615	2643.345	2718.961	NA	5127.607	4852.613

GDP 分季度

指标	国内生产总值_当季值(亿元)
2013 年第 1 季度	128083.5
2013 年第 2 季度	143031.8
2013 年第 3 季度	150719.8
2013 年第 4 季度	166183.6
2014 年第 1 季度	138738
2014 年第 2 季度	155201
2014 年第 3 季度	163467
2014 年第 4 季度	178732.8
2015 年第 1 季度	147961.8
2015 年第 2 季度	166216.4
2015 年第 3 季度	173595.3

analysis.m

```
%% Input these data
filename = 'data.csv';delimiter = ',';startRow = 2;
formatSpec = '%s%s%s%s%s%s%s%s%s%s%[^\n\r]';
fileID = fopen(filename,'r');
dataArray = textscan(fileID, formatSpec, 'Delimiter', delimiter,
'HeaderLines' ,startRow-1, 'ReturnOnError', false);
fclose(fileID);
raw = repmat({''},length(dataArray{1}),length(dataArray)-1);
for col=1:length(dataArray)-1
```

```
    raw(1:length(dataArray{col}),col) = dataArray{col};
end
numericData = NaN(size(dataArray{1},1),size(dataArray,2));
for col=[1,2,3,4,5,6,7,8,9]
    rawData = dataArray{col};
    for row = 1:size(rawData, 1);
        regexstr =
'(?<prefix>.*?)(?<numbers>([-]*(\d+[\,]*)+[\.]{0,1}\d*[eEdD]{0,1}[-+]
*\d*[i]{0,1})|([-]*(\d+[\,]*)*[\.]{1,1}\d+[eEdD]{0,1}[-+]*\d*[i]{0,1}
))(?<suffix>.*)';
       try
           result = regexp(rawData{row}, regexstr, 'names');
           numbers = result.numbers;
           invalidThousandsSeparator = false;
           if any(numbers==',');
               thousandsRegExp = '^\d+?(\,\d{3})*\.{0,1}\d*$';
               if isempty(regexp(thousandsRegExp, ',', 'once'));
                   numbers = NaN;
                   invalidThousandsSeparator = true;
               end
           end
           if ~invalidThousandsSeparator;
               numbers = textscan(strrep(numbers, ',', ''), '%f');
               numericData(row, col) = numbers{1};
               raw{row, col} = numbers{1};
           end
       catch me
       end
    end
end
R = cellfun(@(x) ~isnumeric(x) && ~islogical(x),raw);
raw(R) = {NaN};
T = cell2mat(raw(:, 1));
TAX = cell2mat(raw(:, 2));
GDP = cell2mat(raw(:, 3));
EXP = cell2mat(raw(:, 4));
IE = cell2mat(raw(:, 5));
RS = cell2mat(raw(:, 6));
COM = cell2mat(raw(:, 7));
INV = cell2mat(raw(:, 8));
DEP = cell2mat(raw(:, 9));
clearvars filename delimiter startRow formatSpec fileID dataArray ans raw
clearvars col numericData rawData row regexstr result numbers
invalidThousandsSeparator thousandsRegExp me R;
```

```
%% Regression analysis
y=TAX;x=GDP;
x1=[ones(length(x),1),x];
[b,bint,r,rint,stats]=regress(y,x1);
b,bint,stats,rcoplot(r,rint)
plot(x,y,'o');
f=lsline;
set(f,'Color',[0 0 0])
%% Time series
%TAX
m1=length(y);n=6; %n is the number of moving average
m1
for i=1:m1-n+1
    yhat1(i)=sum(y(i:i+n-1))/n;
end
yhat1
m2=length(yhat1);
for i=1:m2-n+1
    yhat2(i)=sum(yhat1(i:i+n-1))/n;
end
yhat2
a37=2*yhat1(end)-yhat2(end)
b37=2*(yhat1(end)-yhat2(end))/(n-1)
y2015=a37+b37
%GDP
m1=length(x);
for i=1:m1-n+1
    xhat1(i)=sum(x(i:i+n-1))/n;
end
xhat1
m2=length(xhat1);
for i=1:m2-n+1
    xhat2(i)=sum(xhat1(i:i+n-1))/n;
end
xhat2
a37=2*xhat1(end)-xhat2(end);
b37=2*(xhat1(end)-xhat2(end))/(n-1);
x2015=a37+b37
```

figure.R

```
figure <- function()
{
  par(mfrow=c(2,2))
  mydata <- read.csv("data.csv")
```

```
  attach(mydata)
  #Those codes above are data-reading.
  T = 1978:2014
  plot(T,GDP);abline(v=1993,lty=3)
  plot(T,TAX);abline(v=1993,lty=3)
  D = ifelse(T>1993,1,0)
  plot(GDP,TAX)
  abline(h=mydata[T==1993,'TAX'],v=mydata[T==1993,'GDP'],lty=3)
  summary(lm(TAX~GDP+GDP*D))
  plot(TAX~GDP)
  abline(3.1137,0.11787,lty=3);abline(3.1137-79.51279,0.11787+0.08319)
  #Those codes above are mix return.
  return(0)
}
```