

# 全国第六届研究生数学建模竞赛



题 目                      我国就业人数或城镇登记失业率的数学建模

## 摘                      要：

本文是一个研究探讨影响我国城镇就业人数或城镇登记失业率主要因素和指标，建立统计预测模型，并对我国下一时期就业形势作出仿真预测和评价及建议的问题。

首先，在阅读大量文献的基础上确定就业人口数的影响因素，利用国家统计局公布的相关指标的数据，运用主成分分析法选取出影响就业人数的主要指标，如国家财政收入、第三产业产值、社会消费品零售总额、GDP、进出口总额、固定资产投资、外汇储备等等，并用相关系数对该结论进行了验证。接着运用数理统计的知识，对 1996 到 2007 年的数据进行统计回归分析，建立并检验了我国城镇就业人口数与国家财政收入、第三产业产值、社会消费品零售总额等指标之间的统计回归模型（见式 4-2-1），并对模型进行了评价。

接着，基于对实际情况的假设，由浅入深的建立了分地区、分行业、分人群情况下的统计模型（分别见式 4-3-1、式 4-3-4，式 4-3-6），模型的检验结果也较为合理，比较真实的反应了各种情况下就业人口数的变化情况。

然后，依据拟合曲线得出 09 年和 10 年上半年主要影响指标的数值，分别运用统计回归模型（见表 4-4-5）、BP 神经网络模型（见表 4-4-7）和 Logistic 预测模型（见表 4-4-8）对我国城镇就业人数进行仿真预测，借以提高预测的准确性。

最后，结合本文的研究成果和我国的具体国情，在调整模型变量比较变动幅度的基础，对增加我国城镇就业人数提出了中肯建议。

关键词：主成分分析    多元逐步回归模型    BP 神经网络    Logistic 预测模型

参赛队号 1194218

队员姓名 何芬芬 汪博 徐翠翠

参赛密码 \_\_\_\_\_

(由组委会填写)

# 我国就业人数或城镇登记失业率的数学建模

## 1、问题背景

失业问题是一个世界性的问题。我国是一个劳动力相对富余的国家，来自人力资源与社会保障部的数字表明，目前我国平均每年大约有 1400 万左右新增劳动力，加上现存的失业和下岗人员，每年要就业的劳动力达 2400 万人。2009 年我国就业面临更大的挑战，一是国际金融危机导致国际市场需求难以在短期内复苏；二是今年我国经济增长速下滑；三是国内消费需求乏力；四是一些行业产能过剩与市场预期不确定导致企业投资不足，所以就业形势十分严峻。

过量的失业人口引发了社会矛盾，造成了劳动力资源的极大浪费；阻碍了社会生产力的发展；给政府和社会带来了极大的经济负担，成为我国经济社会中最突出的问题。因此对于就业问题的研究和解决极为迫切。

## 2、相关信息

自失业现象产生以来，西方经济学家就对导致失业产生的原因进行了研究，经济学各个流派在研究失业问题的过程中，形成了各具特点、有所侧重的失业理论，并提出了有针对性的政策建议。国内理论界结合中国的改革实践，对原有的马克思主义就业理论进行理论创新，形成了劳动力个人所有制理论、企业是用工主体理论、劳动力成为商品的理论。近年来，国内学者对我国的失业问题进行了更为深入的研究，思路也更为开阔，取得了许多重要成果。现引入这些文献的重要结论，以方便本文对就业影响因素、建模进行分析和研究：

### (1) 政策原因

- 政府宏观调控会影响失业率——不完全劳动市场、政府干预与失业率<sup>[1]</sup>；
- 社会政策目标选择对就业政策的影响<sup>[2]</sup>，而就业政策对就业具有直接影响。——社会政策目标选择对就业政策的影响；
- 充分运用财政、税收、信贷等手段，加大再就业资金投入，这也会影响就业率——转型期中国灵活就业及其制度创新问题研究；
- 对外直接投资对国内就业也有影响<sup>[3]</sup>——中国对外直接投资对国内就业的影响分析；
- 劳动制度因素、就业保障体制对就业的影响。

### (2) 经济原因

- 经济增长影响失业率<sup>[4]</sup>，就业弹性指标<sup>[5]</sup>可以描述经济增长和就业增长之间的关系——吉林省就业弹性变化及影响因素分析；
- 技术进步具有双重的就业效应。经济的增长方式转变也会引起我们就业率的下降和失业率的上升。——我国三次产业就业增长率影响因素分析；
- 如果社会总资本增长速度低于资本有机构成增长速度的条件下，必然引起原有职工下岗，从而导致失业——中国治理失业的对策研究；
- 劳动生产率和就业之间也存在着复杂的联系<sup>[6]</sup>。——中国劳动生产率的就业效应分析；

- 人民币汇率降低, 促进出口, 增加就业; 汇率升高, 出口萎缩, 失业率升高<sup>[7]</sup>;
- 广义货币供应量(M2)、货币政策的最终目标, 一般有四个: 稳定物价、充分就业、促进经济增长和平衡国际收支等;
- 国民产出水平、工资水平、通货膨胀率等都会影响就业率——我国就业率因素的实证研究。

### (3) 结构原因

- 人口构成影响就业结构的合理性会影响就业率。——我国体制转型时期失业的影响因素及就业治理结构;
- 我国的产业结构和就业结构对就业都有较大影响<sup>[8]</sup> ——我国产业结构和就业结构的现状趋势分析以及国际比较;
- 经济结构的快速调整直接影响就业率的高低。

### (4) 社会原因

- 宏观环境, 如经济危机, 所处社会发展阶段等;
- 劳动力市场的供求因素会影响就业率——影响辽宁省就业因素分析;
- 人口数量和质量因素——我国体制转型时期失业的影响因素及就业治理结构;
- 人才培养体制与就业需求的结合程度。

### (5) 个体原因

- 微观角度上, 就业者的自身状况, 如综合素质、学历层次、专业特点、实践能力、性别状况、面试技巧、家庭背景和社会关系等, 则是推动就业的最直观因素。其中, 就业观居于核心地位。

当然还有一些其他影响因素会影响到就业人数, 如宏观经济环境、自然灾害、国家对某行业的扶持政策等, 但这些多属于随机因素, 因此在这里我们不作为决定性因素加以考虑。在阅读文献时我们得知, 影响我国就业率的最主要因素是国内生产总值, 其次是通货膨胀率。我国劳动力市场供大于求和职工工资水平基本相差不大, 所以平均货币工资水平对我国的就业率影响不大。因此本文不考虑平均货币工资水平对就业的影响。

## 3、建模基础

### 3.1 基本假设

- 1、短期内, 我国经济社会建设将一直保持快速发展, 与社会建设、人民物质财富相关的部分因素将满足“J”型曲线增长。
- 2、影响就业的定性因素, 如个体的就业偏好, 价值观取向, 在模型中用常量代替, 并假定其在本文的研究年份不变。
- 3、忽略除本文选定的因素外其他的因素对就业的影响。
- 4、假设近年内不会发生严重的经济危机、自然灾害或者其他不可预见却危害极大的事件, 政治局势稳定。
- 5、假设模型中所涉及的数据都是原始有效, 不考虑特殊情况的存在。
- 6、假设随机误差项具有零均值, 同方差以及序列不相关  $E(\varepsilon_j)=0$ 。

为了选择适当的参数估计方法, 提高估计精度, 通常需要对模型的随机误差

和解释变量的特性事先做些假设。这些假设的作用是便于分离回归模型中每个因素的单独影响，在回归分析的参数估计和统计检验理论中，许多结论都以这些假设为基础。

### 3.2 符号说明

$x_i$ : 表示第  $i$  因素在 1996 年至 2007 年统计量数列，其中  $i=1, 2, 3, \dots, 17$

$x_i(j)$ : 表示第  $i$  因素在第  $j$  年份（1996 年对应第 1 年、依此类推）的数量值，其中  $j=1, 2, 3, \dots$

$y_i$ : 表示第  $i$  年份城镇就业人口数量，其中  $i=1, 2, 3, \dots, 12$

### 3.3 问题分析

本题是关于城镇就业人口数量或者城镇失业率影响因素的甄别、判断，并通过建立就业预测模型预测未来就业形势的综合性社会热点问题。需综合考虑经济、政治、社会、环境、个体等各方面因素，才能建立符合实际的主因素判断模型及预测模型。

问题（1）的解决需要在阅读大量文献的基础上找出影响就业的因素，然后通过主成分分析找出主要就业影响因素。然后关联度方法验证本题的结论。（2）可视为根据查到的因变量和自变量的数据，建立就业人数预测模型，并讨论元素变化对因变量即就业人数影响。由于就业率还会随着行业、地区、就业人群的不同而不同，问题（3）要求进一步细分量化指标，建立起更为精确的数学模型。问题（4）则需要结合国家的财政政策和经济环境，预测 2009 年和 2010 年上半年的就业情况。

建立以上模型，主要用到的方法是数据清理、主成分分析法、相关系数分析、散点图拟合、一元非线性回归分析，多元逐步回归分析、统计数值分析、Logistic 预测模型、BP 神经网络模型等，主要用到的计算机软件有 MATLAB、SPASS 等。

## 4、模型的建立与求解

### 4.1 模型的主要因素和指标

#### 4.1.1 主成分分析确定主要因素

影响就业（或者失业）的因素很多，从宏观层面上，消费、投资、政府购买和进出口都是重要的因素；而从中观层面，不同地区、不同产业也会表现出不同的特征。中央政府调整宏观经济政策（包括财政政策和货币政策），以及对不同地区和不同产业实行不同的扶持政策都会对就业产生巨大的影响。

参照本文相关信息中对就业影响因素的分析、指标数据的可获得性、指标数据的代表性，我们选取了如下指标：

#### A. 政策指标

- （1）固定资产投资（2）进出口总额（3）广义货币供应量（M2）（4）外汇储备（5）国家财政收入（6）国家财政支出

#### B. 经济指标

- （7）CPI（8）PPI（9）GDP（10）通货膨胀率（11）我国全社会消费品零售总额（12）、人民币汇率（13）技术进步（劳动生产率）

C. 结构指标

(14) 三次产业增加值

D. 社会指标

(15) 宏观环境 (16) 劳动力的供给 (17) 社会就业需求

E. 个体指标

(18) 劳动者的素质、价值观

说明:

本指标的选取,除了个体指标外定性指标几乎没有考虑,为方便建模,本文假定定性指标在短时间内不会有大幅变动,在模型中用一个常量代替。

然后用数据清理的方法,根据中国统计局官方网站公布的数据制作了就业影响因素表格,如表 4-1-1 所示:

表 4-1-1 1996——2007 年我国城镇就业人口及其影响指标的统计

年份	城镇就业人口数(万) (y)	CPI (x <sub>1</sub> )	GDP(亿) (x <sub>2</sub> )	固定资产投资 (亿) (x <sub>3</sub> )	进出口总额 (亿) (x <sub>4</sub> )	广义货币供应量 (M2) (x <sub>5</sub> )	服务业增长值 指数 (x <sub>6</sub> )	人民币平均汇率(对美元) (x <sub>7</sub> )	国家财政收入 (亿) (x <sub>8</sub> )
1996	19922.00	108.80	71176.59	22913.55	24133.8	0.10	109.4265	831.42	7407.99
1997	20781.00	103.10	78973.03	24941.11	26967.2	0.09	110.717	828.98	8651.14
1998	21616.00	99.40	84402.28	28406.3	26849.7	0.08	108.3723	827.91	9875.95
1999	22412.00	98.70	89677.05	29855.76	29896.2	0.06	109.3298	827.83	11444.08
2000	23151.00	100.80	99214.55	32917.73	42183.6	0.06	109.7489	827.84	13395.23
2001	23940.00	100.70	109655.2	37213.49	42183.6	0.07	110.2574	827.7	16386.04
2002	24780.00	99.00	120332.7	43499.91	51378.2	0.07	110.4423	827.7	18903.64
2003	25639.00	100.90	135822.8	55566.61	70483.5	0.07	109.503	827.7	21715.25
2004	26476.00	103.30	159878.3	70477.4	95539.1	0.08	110.0564	827.68	26396.47
2005	27331.00	101.60	159878.3	88773.6	116921.8	0.08	110.5002	819.17	31649.29
2006	28310.00	101.50	211923.5	109998.2	140971.4	0.08	112.1064	797.18	38760.2
2007	29350.00	104.50	249529.9	137323.9	166740.2	0.10	112.5525	760.4	51321.78

续表 4-1-1

年份	通货膨胀率 ( $x_9$ )	全社会消费品零售总额(亿) ( $x_{10}$ )	第一产业增加值 (亿) ( $x_{11}$ )	第二产业增加值 (亿) ( $x_{12}$ )	第三产业增加值 (亿) ( $x_{13}$ )	财政支出(亿) ( $x_{14}$ )	文教、科学、卫生支出 (亿) ( $x_{15}$ )	外汇储备(10亿美元) ( $x_{16}$ )	PPI ( $x_{17}$ )
1996	8.3	28360.2	19.5	47.5	33	7937.55	1704.25	105.049	4.5
1997	2.8	31252.9	18.1	47.5	34.4	9233.56	1903.59	139.89	2.8
1998	-0.8	33378.1	17.3	46.2	36.5	10798.18	2154.38	144.959	-0.8
1999	-1.4	35647.9	16.2	45.8	38	13187.67	2408.06	154.675	-1.4
2000	0.4	39105.7	14.8	45.9	39.3	15886.50	2736.88	165.574	0.4
2001	0.7	43055.4	14.1	45.2	40.7	18902.58	3361.02	212.165	0.7
2002	-0.8	48135.9	13.5	44.8	41.7	22053.15	3979.08	286.407	-0.8
2003	1.2	52516.3	12.5	46	41.5	24649.95	4505.51	403.251	1.2
2004	3.9	59501.0	13.1	46.2	40.7	28486.89	5143.65	699.32	3.9
2005	1.8	67176.6	12.5	47.5	40	33930.28	6104.18	818.872	1.8
2006	1.5	76410.0	11.3	48.7	40	40422.73	7425.98	1066.3	1.5
2007	4.8	89210.0	11.1	48.5	40.4	49781.35	11793.96	1527.9	2.2

城镇就业人口数与表 4-1-1 中列出的这些因素具有一定联系,但并非每一指标都对城镇就业人口数产生显著影响。指数太多,无疑会增加分析问题的难度与复杂性,而且这多个指标之间是可能具有一定的相关关系的。此时主成分分析法可以用来确定影响就业的主要因素或指标。

主成分分析(Principal Component Analysis)<sup>[9]</sup>是一种在处理多个变量时经常使用的一种方法。主成分分析法可以将彼此相关的指标变量转化为彼此不相关的指标变量;将个数较多的指标变量转化为个数较少的指标变量;将意义单一的指标变量转化为意义综合的指标变量。将彼此相关的变量转变为彼此不相关的新变量、方差较大的几个新变量就能综合反应原多个变量所包含的主要信息;因此就可以找出主要的影响因素。

主成分分析计算步骤如下:

(1)计算相关系数矩阵

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ M & M & M & M \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \quad (4-1-1)$$

在公式 (4-1-1)中,  $r_{ij}$  ( $i, j=1, 2, \cdots, p$ )为原来变量  $x_i$  与  $x_j$  的相关系数, 其计算公式为

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}} \quad (4-1-2)$$

因为  $R$  是实对称矩阵 (即  $r_{ij}=r_{ji}$ ), 所以只需计算其上三角元素或下三角元素即可。

原始数据的标准化处理。首先将表 4-1-1 中的原始数据作标准化处理, 由公式 (2) 计算得相关系数矩阵。

(2) 计算特征值与特征向量

首先解特征方程  $|\lambda I - R| = 0$  求出特征值  $\lambda_i$  ( $i=1, 2, \cdots, p$ ), 并使其按大小顺序排列, 即  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ ; 然后分别求出对应于特征值  $\lambda_i$  的特征向量  $e_i$  ( $i=1, 2, \cdots, p$ )。由相关系数矩阵计算特征值, 以及各个主成分的贡献率与累计贡献率。由表 4-1-2 可知, 第一, 第二主成分的累计贡献率已高达 90.636%, 故只需求出第一, 第二主成分  $z_1, z_2$  即可。计算主成分贡献率及累计贡献率公式:

主成分  $z_i$  贡献率:  $r_i / \sum_{k=1}^p \gamma_k$  ( $i=1, 2, \cdots, p$ ), 累计贡献率:  $\sum_k^m \gamma_k / \sum_k^p \gamma_k$ 。

表 4-1-2 特征值及主成分贡献率

主成分	特征值	方差贡献率	累计贡献率
1	10.970	64.530	64.530
2	4.438	26.106	90.636
3	0.851	5.006	95.642
4	0.279	1.642	97.284
5	0.276	1.624	98.908
6	0.114	0.668	99.576
7	0.038	0.226	99.801
8	0.024	0.140	99.941
9	0.008	0.048	99.989
10	0.001	0.008	99.997
11	0.000	0.003	100.000
12	0.000	0.000	100.000

13	0.000	0.000	100.000
14	0.000	0.000	100.000
15	0.000	0.000	100.000
16	0.000	0.000	100.000
17	0.000	0.000	100.000

### (3) 计算主成分载荷

$$p(z_k, x_i) = \sqrt{\gamma_k} e_{ki} (i, k=1, 2, \dots, p) \quad (4-1-3)$$

由此可以进一步计算主成分得分：

$$z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ M & M & M & M \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix} \quad (4-1-4)$$

对于特征值  $\lambda_1=10.970$ ， $\lambda_2=4.438$  分别求出其特征向量  $e_1$ ， $e_2$ ，并计算各变量  $x_1$ ， $x_2$ ， $\dots$ ， $x_{17}$  在各主成分上的载荷得到主成分载荷矩阵。

以上都是步骤都是应用 SPSS 统计分析软件进行主成分分析，求出相关矩阵的特征根  $\lambda$ 、特征向量  $\mu$  及累计方差贡献率  $E$  及主成分载荷矩阵的。

表 4-1-3 主成分载荷矩阵

指标	主成分		指标	主成分	
	$Z_1$	$Z_2$		$Z_1$	$Z_2$
$x_1$	0.204	0.834	$x_{10}$	0.989	-0.115
$x_2$	0.989	-0.115	$x_{11}$	0.994	-0.042
$x_3$	0.994	-0.042	$x_{12}$	0.988	-0.049
$x_4$	0.988	-0.049	$x_{13}$	0.315	-0.876
$x_5$	0.315	0.876	$x_{14}$	0.858	0.087
$x_6$	0.858	0.087	$x_{15}$	-0.895	-0.080
$x_7$	-0.895	-0.080	$x_{16}$	0.993	-0.104
$x_8$	0.993	-0.104	$x_{17}$	0.276	0.902
$x_9$	0.276	0.902			

从表 4-1-3 可以看出，第一主成分  $z_1$  与  $x_2$ ， $x_3$ ， $x_4$ ， $x_8$ ， $x_{10}$ ， $x_{11}$ ， $x_{13}$ ， $x_{14}$ ， $x_{15}$  有较大的正相关；第二主成分  $z_2$  与  $x_{13}$  有较大的负相关。

以上分析结果表明，根据主成分贡献率，用两个主成分可以代替这 16 个指标，大大降低了计算复杂度。根据主成分载荷矩阵可以看到， $x_2$ ， $x_3$ ， $x_4$ ， $x_8$ ，



$X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{16}$  对第一主成分  $z_1$  影响很大,  $X_{12}$  对第二主成分  $z_2$  影响很大, 所以综合影响这两个主成分指标。

可得出影响就业的主要指标如下: GDP、固定资产投资、进出口总额、国家财政收入、全社会消费品总额、财政支出、第一产业、第三产业、文教科学卫生支出、外汇储备。

#### 4.1.2 模型验证和评价

为了进一步验证所选的主要影响因素满足相关性最大的原则。利用皮尔逊相关系数来分析系统中就业人口数与各因素关联程度。

皮尔逊相关系数又称“皮尔逊积矩相关系数”<sup>[10]</sup>, 对两个定距变量(例如, 年龄和身高)的关系强度的测量, 简写  $\tau$ 。这一测量也可用作对显著性的一种检验, 其方法是检验解消假设: 总体中的  $\tau$  值为 0。若样本  $\tau$  实际上不等于 0, 则解消假设可加否定, 从而我们可以满意地看到, 这两个变量不是无关的, 在统计显著性层次上它们是有关的。例如, 若我们有一个较大的样本, 并发现一个高的样本值  $\tau$  (例如, 90), 那么我们不妨否定这一解消假设: 这个样本是来自一个其真正的  $\tau$  值为 0 的总体, 因为假若真正的总体值是 0, 我们就不可能单纯碰巧取得一个如此高的样本。 $\tau$  的变化从 -1 (全负关系), 通过 0 (无关系或无关性), 到 +1 (全正关系)。从直线关系和曲线关系之间的关系来说,  $\tau$  是对直线关系的一种测量。对  $\tau$  有两个主要的解释: (1)  $\tau^2$  = 所解释的方差额。(2)  $\tau$  测量围绕回归线散布的程度, 也就是说, 它告诉我们, 我们用回归线可预测的准确程度有多大。

模型评价: 由上表可以看到, 与  $y$  相关系数大于 0.9 的有  $X_2, X_3, X_4, X_8, X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{16}$  与用主成分分析法求得的数据只差了一个  $X_{13}$ , 其它数据指标都吻合, 说明用主成分分析法得出影响我国城镇就业人口的主要指标的方法还是比较准确的。

## 4.2 城镇就业人口的统计回归模型

### 4.2.1 问题分析

要探讨城镇就业人口数, 作为建立新的模型前的准备, 首先要了解长期以来我国城镇就业人口数的变化与规律。通过研究分析第一小题中求出的影响城镇就业人口数的影响因素, 利用统计数学的逐步回归分析方法, 可以得出影响城镇就业人口数的多因素统计回归模型, 这就为之后的新标准探讨提供了相关数据及对比依据。

逐步回归是一种从众多自变量中有效的选择重要变量的方法, 其基本思路就是先确定一个包含若干自变量的初始集合, 然后每次从集合外的变量中引入一个对因变量影响最大的, 再对集合中的变量经行检验, 从变得不显著的变量中移出一个影响最小的, 依次进行, 直到不能引入和移出为止, 引入和移出都是以给定的显著性水平为标准。

表4-2-1 1996年到2007年城镇就业人口与影响因素

## 4.2.2 多元逐步回归模型

首先，为确定各单因素与均生学费间关系，可利用表4-2-1分别做出城镇就业人数 $y$ 对 $x_1$ 、 $x_2$ 、 $x_3$ 、 $x_4$ 、 $x_5$ 、 $x_6$ 、 $x_7$ 、 $x_8$ 、 $x_9$ 、 $x_{10}$ 的散点图并选用恰当线型进行曲线拟合。

由预处理表中的相关信息，利用 MATLAB 软件，首先做出就业人数与各单因素的散点图，画出理想拟合曲线，国家财政收入的图例如下：

年份	城镇 就业 人口 数 (万) ( $y$ )	GDP(亿) ( $x_1$ )	固定资产 投资(亿) ( $x_2$ )	进出口总 额 ( $x_3$ )	国家财政 收入(亿 元) ( $x_4$ )	全社会 消费品 零售总 额(亿 元) ( $x_5$ )	第一 产业 ( $x_6$ )	第三 产业 ( $x_7$ )	财政支出 (亿元) ( $x_8$ )	文教、科 学、卫生 支出(亿 元) ( $x_9$ )	外汇储 备(10 亿美元) ( $x_{10}$ )
1996	19922	71176.59	22913.55	24133.8	7407.99	28360.2	19.5	33	7937.55	1704.25	105.049
1997	20781	78973.03	24941.11	26967.2	8651.14	31252.9	18.1	34.4	9233.56	1903.59	139.89
1998	21616	84402.28	28406.3	26849.7	9875.95	33378.1	17.3	36.5	10798.18	2154.38	144.959
1999	22412	89677.05	29855.76	29896.2	11444.08	35647.9	16.2	38	13187.67	2408.06	154.675
2000	23151	99214.55	32917.73	42183.6	13395.23	39105.7	14.8	39.3	15886.50	2736.88	165.574
2001	23940	109655.2	37213.49	42183.6	16386.04	43055.4	14.1	40.7	18902.58	3361.02	212.165
2002	24780	120332.7	43499.91	51378.2	18903.64	48135.9	13.5	41.7	22053.15	3979.08	286.407
2003	25639	135822.8	55566.61	70483.5	21715.25	52516.3	12.5	41.5	24649.95	4505.51	403.251
2004	26476	159878.3	70477.4	95539.1	26396.47	59501.0	13.1	40.7	28486.89	5143.65	699.32
2005	27331	159878.3	88773.6	116921.8	31649.29	67176.6	12.5	40	33930.28	6104.18	818.872
2006	28310	211923.5	109998.2	140971.4	38760.2	76410.0	11.3	40	40422.73	7425.98	1066.3
2007	29350	249529.9	137323.9	166740.2	51321.78	89210.0	11.1	40.4	49781.35	11793.96	1527.9

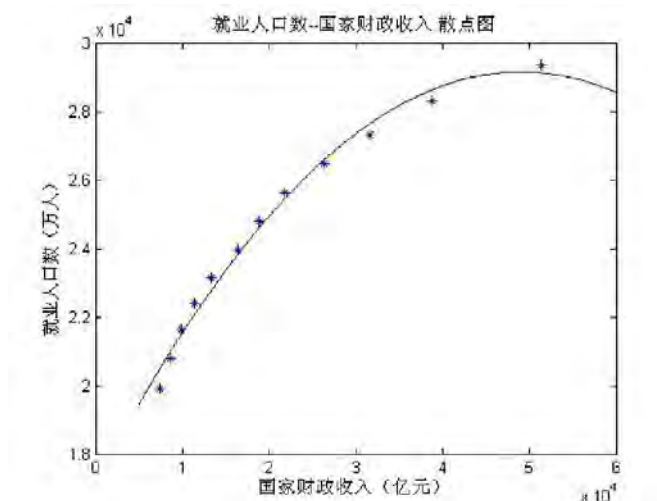


图 4-2-1 国家财政收入散点图

对此问题的分析由简入繁，先假设整个模型呈线性模型，若拟合度较差，则引入二次模型，三次模型，对数模型等等。

MATLAB 统计工具箱中的逐步回归命令是 `stepwise`<sup>[11]</sup>，它提供人机交互式画面，研究者可以在画面上自由地引入和移出变量，进行统计分析。

`stepwise(x, y, 5, 0.05)` 命令下,  $x$  是自变量矩阵 ( $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$ )，首先引入一次模型  $Y = \beta X + \varepsilon$ ，接着引入二次模型、三次模型、对数模型等等，发现效果经过实验发现一次模型的效果最好，所以选择一次模型。首次设定所有变量进入初始集合，显著性水平为 0.05。经分析最终选择保留  $x_4, x_5, x_7$  作为检验变量。

利用MATLAB统计工具箱中的命令`regress`求解，可得到此多因素回归分析模型参数估计值及其置信水平（置信水平  $\alpha = 0.05$ ）、检验统计量  $R^2$ 、 $F$ 、 $p$  的结果，得到如下逐步回归统计分析模型：

$$y = 6176.156520 - 0.2187870x_4 + 0.285193387x_5 + 221.784374x_7 \quad (\text{式4-2-1})$$

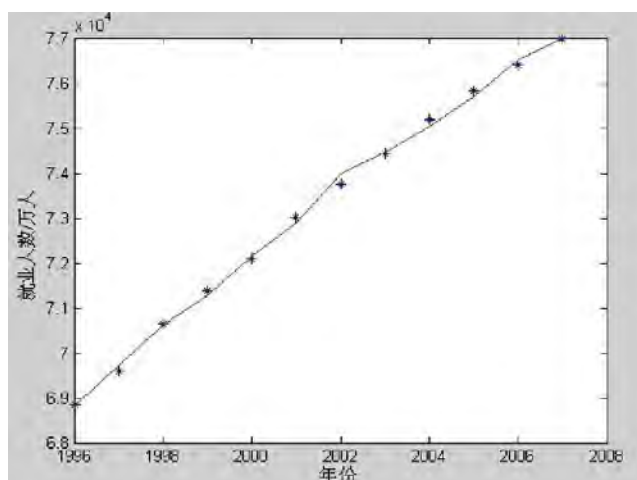


图4-2-2 1996年至2007年城镇就业人口变化与预测

### 4.2.3 模型分析与检验

#### (1) 残差分析

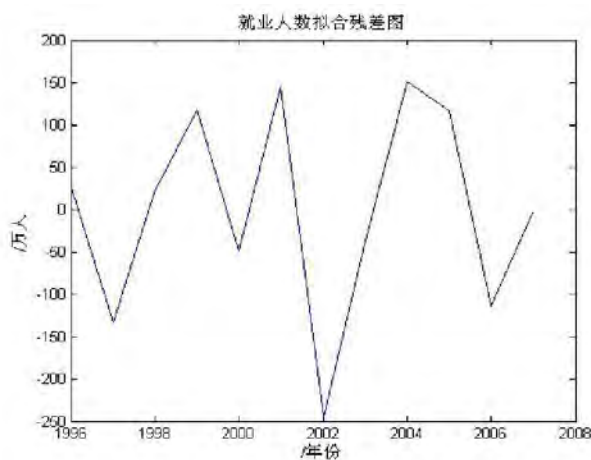


图 4-2-3 城镇就业人数拟合残差图

表 4-2-2 预测值与残差值

年份	预测值 (万人)	真实值 (万人)	残差 ( $y - \hat{\beta}x$ )
1996	19962.544	19922	-40.544
1997	20826.0416	20781	-45.0416
1998	21629.9177	21616	-13.9177
1999	22266.84578	22412	145.1542
2000	23114.42606	23151	36.57394
2001	23897.00764	23940	42.99236
2002	25016.90276	24780	-236.90276
2003	25606.66228	25639	32.33772
2004	26397.03138	26476	78.96862
2005	27281.56075	27331	49.43925
2006	28359.09023	28310	-49.09023
2007	29349.96970	29350	0.0303

由上图及表 4-2-2 可以得出，残差值基本围绕 0 上面波动，最大振幅不超过 237（万人），较为理想。因此可以说明，模型对 1996 年—2007 年就业人数的预测效果良好，基本可以接受的。

#### (2) 拟合优度检验

$R = 0.999045$ ，一般情况下只要大于 0.9 就算通过检验。故此模型可以接受的。

#### (3) 方程显著性检验

$F = 2789.7569$ ， $F$  临界值为 0.0001， $F$  值均远远超过  $F$  检验的临界值，故可通过方程显著性检验。

#### (4) 变量显著性检验

表 4-2-3 参数估计值及其置信区间

参数	参数估计值	参数置信区间
$\beta_0$	6176.15652063512	4960.70011799588, 7391.61292327435
$\beta_1$	-0.218787067827949	-0.294311588572612, -0.143262547083286
$\beta_2$	0.285193387198799	0.229927372405107, 0.340459401992491
$\beta_3$	221.788437419509	178.712630727679, 264.864244111338
$R^2 = 0.999045$ $F = 2789.7569$ $p < 0.204591898977924e-11$ $F_{0.05}(11,3) = 12047.5991244$		

表 4-2-3 上给出了各检验变量参数回归系数值  $\beta_1$ 、 $\beta_2$ 、 $\beta_3$ 、 $\beta_4$  及其置信区间，检查他们置信区间发现均不包含零点，在此表明， $x_4$ 、 $x_5$ 、 $x_7$  对因变量 Y 值的影响均是显著的，是满足变量的显著性要求的。

#### 4.2.4 模型总结

此模型均较为理想的通过了以上检验，说明模型是可用的，在一定程度上反映了 1996 年至 2007 年间我国城镇就业人口数变化规律和与相关因素的内在联系。

$$y = 6176.156520 - 0.2187870x_4 + 0.285193387x_5 + 221.784374x_7$$

从拟合表达式上看， $x_4$ 代表的是国家财政收入，其系数值为-0.2187870，相比其它系数值， $x_4$ 所代表的财政收入显然不应与我国城镇就业人数成负相关，这是拟合表达式与现实不太相符的一种体现。 $x_5$ 为全社会消费品零售总额，对就业人数起着正面影响作用，这与现实是相符合的，全社会消费品零售总额越大，城镇就业人口数越多。 $x_7$ 为第三产业，对就业人数的影响最大，这是因为在现在的社会环境下，第三产业可以为社会提供更多的就业岗位。

### 4.3 分行业、地区、就业人群建立精确的数学模型

#### 4.3.1 模型一：分地区建模

首先从中国国家统计局网上查找到影响就业人口数量的主要经济社会指标。包括以下一些指标：地区生产总值（亿元）、第一产业、第三产业、人均地区生产总值（元）、固定资产投资总额（亿元）、房地产开发 地方财政收入（亿元）、地方财政支出（亿元）、货物进出口总额（亿美元）、居民消费价格总指数（上年=100）、社会消费品零售总额（亿元）、毕业生数（万人）、卫生机构数（个）、城镇居民可支配收入

根据主要影响因素的相关性，选取表 4-3-1 中的四个主要影响因素。

表 4-3-1 2001 年至 2007 年西部地区城镇就业人数与影响因素数据

西部地区 就业数据	城镇就业人 数(万人)	第三产业 (亿元)	固定资产投资 总额(亿元)	城镇居民可支 配收入(元)	居民消费价格总 指数(上年=100)
2001	3462	6984.8	7158.8	6169.9	101.5
2002	3462	7760.4	8515.4	6675.2	99.8
2003	3628	8668.2	10843.5	7205	101.4
2004	3710	9986.4	13754.4	8031	103.6
2005	3896	13237.1	17645	8783	101.4
2006	4036	15251.4	21996.9	9728	101.7
2007	4340.4	9850.3	28250.9	11309	105.4

建模准备：假设影响每地区就业人口数的因素和比重是基本相同的。

建模过程：

(1) 由于统计局统计为了涵盖尽量多的信息，所以数据信息的冗余度较大。为了分析其他影响因素没有进入最终模型的原因，可以计算个因素的相关系数。利用 MATLAB 统计工具箱中 corrcoef 命令直接得到各因素的相关系数矩阵。一般认为两个变量的相关系数超过 0.85 时才有显著的相关关系。

通过对前面所收集数据（附 excel 表）的初步分析，主要选取四个主要显著性因素，分别为第三产业、固定资产投资总额、城镇居民可支配收入和居民消费价格总指数。

(2) 基于假设，可由西部地区相关数据来分析建模，以四个显著性因素为因子进行拟合。利用 matlab 统计工具箱中的命令 regress 求解。得出以下方程式：

$$y = 0.151 * x_1 + 0.07306 * x_2 + 0.764 * x_3 - 0.000007195 * x_1^2 + 0.000004663 * x_2^2 - 0.00009147 * x_3^2 + 0.074248 * x_4^2 \quad (\text{式 4-3-1})$$

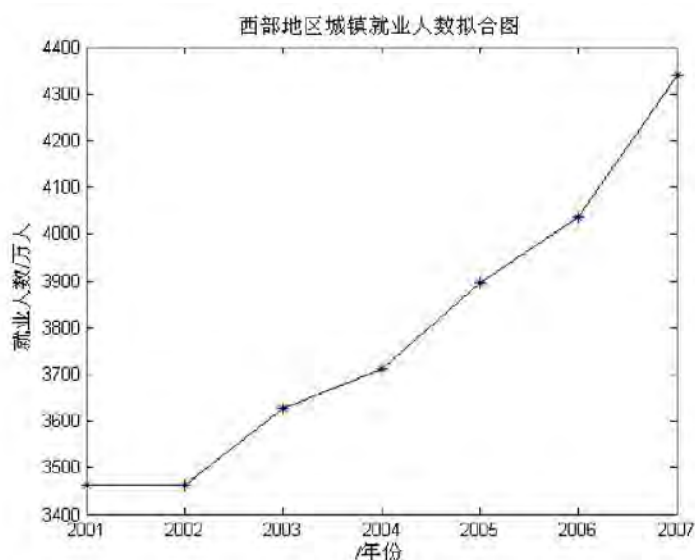


图 4-3-2 2001 年至 2007 年西部城镇就业人数变化与预测

(3) 残差分析：

表 4-3-2 残差值及比例表

年份	预测值（万人）	实际值（万人）	残差
2001	3462	3462	-0.3637978807e-11
2002	3462	3462	0.9094947018e-12
2003	3628	3628	0
2004	3710	3710	-0.4547473509e-12
2005	3896	3896	-0.9094947018e-12
2006	4036	4036	-0.1364242053e-11
2007	4340.4	4340.4	-0.5456968211e-11

从表 4-3-2 中可以看出残差值非常小，完全可以忽略。所以这个模型的预测效果非常好，完全可以接受。

(4) 因此可以得出地区划分下的就业人数统计模型。由于假设的存在，认为结果满足以  $\mu$  为平均数，以  $5\mu$  为方差的正态分布。

(5) 模型检验，以东北地区数据为例。

表 4-3-3 2001 年至 2007 年东北地区城镇就业人数与影响因素数据

东北地区数据	职工人数（万人）	第三产业	固定资产投资总额（亿元）	城镇居民可支配收入	居民消费价格指数（上年=100）
2001	1985.3	3943.6	3086.5	5521.1	100.9
2002	1985.1	4345.8	3485.9	6295.1	99.2
2003	1907	4776.9	4211.6	6981	101.3
2004	1992	5401.7	5579.5	7775	103.8
2005	1956	6442.4	7678.8	8730	101.3
2006	1195.1	7318.4	10520	9830	101.5
2007	1187.3	8516.5	13920.1	11463	105.1

将各因素的数据带入式（4-3-1），通过对计算数据的分析，对拟合方程加上一个常数 200 可以达到良好的拟合效果。得到以下的拟合方程式：

$$y = 200 + 0.151 * x_1 + 0.07306 * x_2 + 0.764 * x_3 - 0.000007195 * x_1^2 + 0.000004663 * x_2^2 - 0.00009147 * x_3^2 + 0.074248 * x_4^2 \quad (\text{式 4-3-2})$$

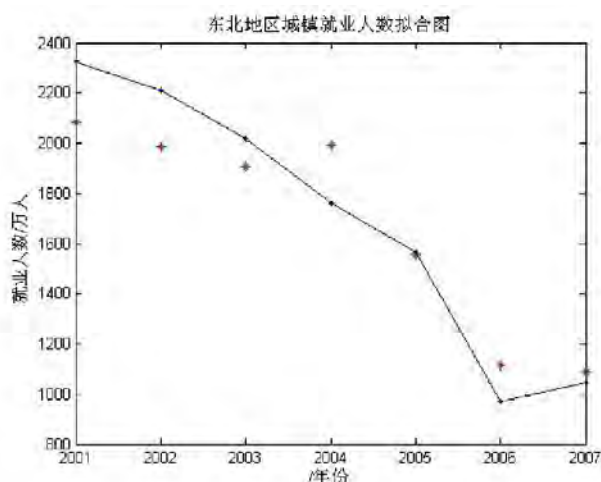


图 4-3-3 2001 年至 2007 年东北地区城镇就业人数变化与预测  
(6) 建模结果分析

表 4-3-4 东北地区城镇就业人数预测值与残差值

模型值	原值	残差值
2325.773	2085.3	240.4731
2210.994	1985.1	225.8935
2017.283	1907	110.283
1762.261	1992	-229.739
1563.333	1556	7.332545
968.5567	1115	-146.443
1045.261	1087	-41.7388

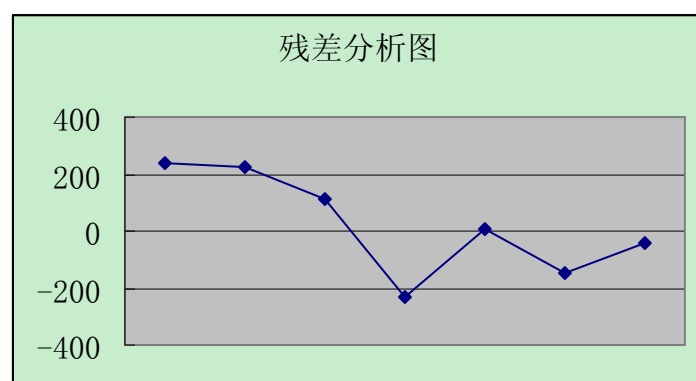


图 4-3-4 残差分析图

由图 4-3-4 及表 4-3-4 可以得出，残差值基本围绕 0 波动，最大振幅在 200 万人左右，对于受多因素综合影响的因变量就业人数来说，较为理想。因此可以说明，模型对 2001 年—2007 年西部地区就业人口数的预测效果良好，基本可以接受的。

通过对 4-3-3 的分析，可以验证西部地区求得的就业人口数学模型能在东部地区得到较好的拟合。同样，对于其它待求的不同地区的就业人数，只要知道这四个显著影响因素的具体值就可以得出当地的就业人口数。如果对公式中的常数进一步细化，可以完全不用考虑地区的划分标准了，只需要关联到该地区的四个影响因素就可以了。

综上所述，可以得出在划分不同地区的情况下，就业人口数在显著影响因素第三产业产值、固定资产投资总额、城镇居民可支配收入、居民消费价格总指数四者影响下的数学模型：

$$y_1 = 200 + 0.15 * x_1 + 0.0731 * x_2 + 0.764 * x_3 - 0.00000719 * x_1^2 + 0.00000466 x_2^2 - 0.00009147 * x_3^2 + 0.074248 * x_4^2 \quad (\text{式 4-3-3})$$

此模型较为真实的反应了在地区划分情况下的就业人口数在不同因素影响下的变化，加入对行业的划分，就可以建立某一地区不同行业的就业人数统计模型。

#### 4.3.2 模型二：任意地区划分行业情况下的数学模型



按照模型一，建立行业划分下的就业人数模型：

表 4-3-5 2000 年至 2007 年航空业从业人数及影响因素数据

	航空从业人数 (万人)	民用航空客运量 总计 (万人)	民用航空航线里 程 (万公里)	民用航班飞行 机场 (个)	民用飞机 架数 (架)
	$y_2$	$x_1$	$x_2$	$x_3$	$x_4$
2000	11.7	6722	150.29	139	982
2001	12.1	7524	155.36	143	1031
2002	12.1	8594	163.77	141	1112
2003	17.9	8759	174.95	126	1160
2004	20.5	12123	204.49	133	1245
2005	21	13827	199.85	135	1386
2006	22.5	15968	211.35	142	1614
2007	23.117	18576	234.3	148	1813

根据航空业的相关数据来分析建立数学模型，利用 matlab 统计工具箱中的命令 regress 求解，以四个显著性因素为因子拟合得出以下统计回归模型：

$$y_2 = 0.00228 * x_1 + 7.761 * x_2 - 4.263 * x_3 - 0.605 * x_1^2 + 0.000000167 x_2^2 + 0.01491 * x_3^2 + 0.000195 * x_4^2 \quad (\text{式 4-3-4})$$

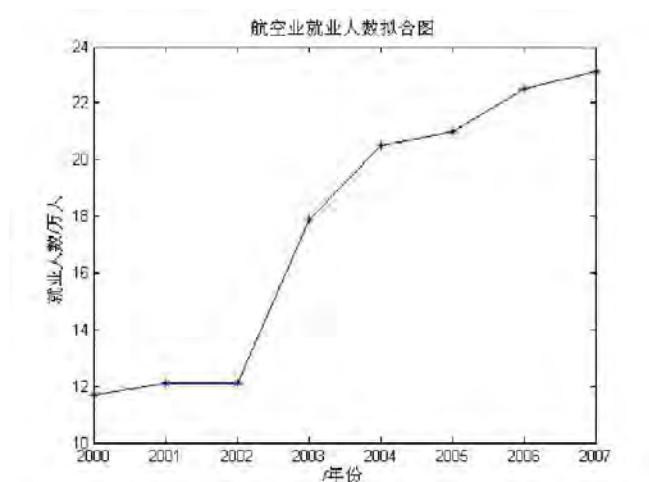


图 4-3-5 航空业就业人数统计模型结果分析图

该模型与模型一中的西部地区的模型相类似，残差值极小，在统计的数据范围内完全可以忽略。所以对航空业的就业人数我们得到一个拟合度很好的模型。

模型检验：根据地区划分和行业划分的原则，查找到一组西部地区航空业的数据，见表 4-3-6。

表 4-3-6 西部某地航空业从业人数与相关数据

年份	航空从业 人数(万 人)	民用航空客运量 总计 (万人)	民用航空航 线里程 (万 公里)	民用航班飞 行机场 (个)	民用飞机架数 (架)
2007	1.21	594	15.5	14	92

将这组数据带入模型二的式 4-3-4 中,可以得到预测人数为 1.26 万人,残差值为 0.05 万人。误差控制了 4.2% 以内,是可以接受的,也说明按照行业划分的模型是具有现实意义的。

综上所述,根据航空业显著性因素民用航空客运量总计、民用航空航线里程、民用航班飞行机场、民用飞机架数所建立的数学模型是可以接受的,且此模型较为真实的反应了在行业划分情况下的就业人口数在不同因素影响下的变化,可普遍用于地区划分下的航空业就业人口预测。

#### 4.3.3 模型三：分就业人群建模

根据模型二,在不同地区及行业划分情况下的就业人群划分,需要做进一步的假设,才能建立最终的符合要求精细程度的模型。但由于中国国家统计局官方网站上暂时不能查找到具体就业人群的划分数据,根据相关介绍和个人经验选择待进入此行业人数、已进入此行业人数、该行业 GDP 和该行业产品价格指数为显著性因素。现假设数据如表 4-3-7:

表 4-3-7 某地区某行业就业人群数据统计

某地区某行业数据统计	就业人群 $M_1$	就业人群 $M_2$	.....	就业人群 $M_n$
待进入此行业人数	$M_{11}$	$M_{21}$	.....	$M_{n1}$
已进入此行业人数	$M_{12}$	$M_{22}$	.....	$M_{n2}$
该行业 GDP	$M_{13}$	$M_{23}$	.....	$M_{n3}$
该行业产品价格指数	$M_{13}$	$M_{24}$	.....	$M_{n4}$

运用模型二中的建模方法,可以建立统计分析模型,其就业人口数可表示为:

$$y_3 = F(M_1, M_2, M_3, \dots, M_n) \quad (\text{式 4-3-5})$$

因此对于某地区某行业的就业人群  $M_i$ , 其就业人口数就可表示为:

$$y_{3i} = F(M_{1i}, M_{2i}, M_{3i}, \dots, M_{ni}) \quad (\text{式 4-3-6})$$

模型分析：由于就业人群划分方法较难得到详实可靠的数据，所以我们采用模型假设的方法，根据模型二的建立模式，建立一个就业人群划分下的就业人口统计分析模型。该模型可以普遍用于地区和行业划分情况下的某就业人群的就业人数预测。

4.4 模型仿真和预测

4.4.1 问题分析

对于二中建立的模型，可预测我国未来的就业人口数。预测值为  $\hat{y}$ ，得到模型的预测方程

$$\hat{y} = 6176.156520 - 0.2187870x_4 + 0.285193387x_5 + 221.784374x_7 \quad (\text{式 } 4-4-1)$$

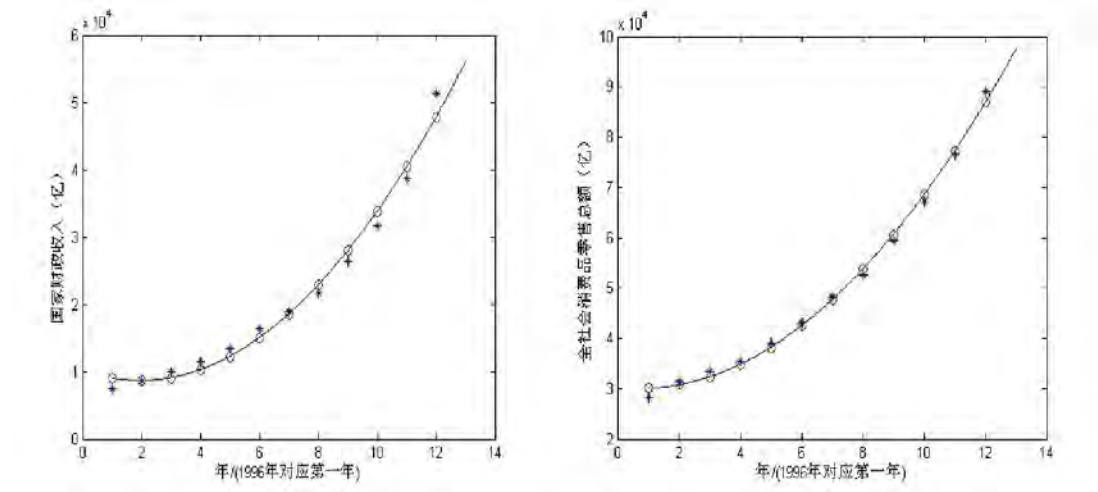
只需知道该未来两年内的国家财政收入，全国消费品零售总额和第三产业

值，就可以计算出预测值  $\hat{y}$ 。

也可以通过求出城镇就业人数的三大因素的时间序列拟合函数，代入到预测模型中，既可以得到一个关于时间t的函数模型。根据时间就可以求得2009和2010年的城镇就业人口。

4.4.2 建模与求解

由问题附表数据预处理表中相关信息，利用 MATLAB 软件，首先做出各单因素与年份的散点图，画出理想拟合曲线，国家财政收入  $x_4$ ，全社会消费品零售总额  $x_5$ ，第三产业  $x_7$ ，作图拟合结果如图 4-4-1。



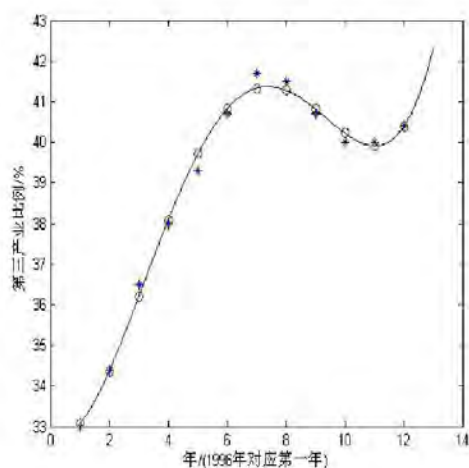


图 4-4-1 各单因素与年份的散点图

由各单因素与年份的散点图分析结果可知：

国家财政收入  $x_4$ ，全社会消费品零售总额  $x_5$  两个因素均适宜用二次曲线进行拟合<sup>[12]</sup>，拟合函数为：

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

利用 MATLAB 统计工具箱中的命令 regress 求解，可得到各单因素回归分析模型的参数估计值及其置信水平（置信水平  $\alpha=0.05$ ）、检验统计量  $R^2$ 、 $F$ 、 $p$ 、 $f$  的结果。

表4-4-1 一元二次回归分析的时间序列拟合参数表

	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_0$ 置信区间	$\beta_1$ 置信区间	$\beta_2$ 置信区间	$R^2$	$F$	$p$	$f$
$x_4$	10221.51	-1539.08	389.69	[5883.51, 14559.41]	[-3073.31, -4.84]	[274.80, 504.58]	0.98	287.79	0.69e-8	3442536.40
$x_5$	30323.38	-654.82	447.61	[27122.42, 33524.35]	[-1786.93, -477.29]	[362.83, 532.38]	0.99	1088.57	0.18e-10	1874438.01

由各因素的时间序列曲线拟合及各参数表可以得出：人国家财政收入  $x_4$ ，

全社会消费品零售总额  $x_5$  的回归方程的决定系数  $R^2$  在 0.99 左右，可靠度极高；

且  $\beta$  置信区间均不包含零点，各变量对因变量影响显著； $F$  值均远远超过  $F$  检验的临界值； $P$  远小于  $\alpha$ 。

所以可以得到：

$$x_4=10221.5 - 1539t + 389.69t^2 \tag{式 4-4-2}$$

$$x_5=30323.32 - 654.82t + 447.6t^2 \tag{式 4-4-3}$$

第三产业的模型采用四次曲线来拟合，，拟合函数为：

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \beta_4x^4$$

表 4-4-2 一元四次回归分析的时间序列拟合参数表

	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_0$ 置信区间	$\beta_1$ 置信区间	$\beta_2$ 置信区 间	$\beta_3$ 置信区 间	$\beta_4$ 置信区 间
$x_7$	32.91	-0.63	0.92	-0.13	0.52e- 2	[31.20, 34. 63]	[-2.30, 1.0 4]	[0.42, 1 .41]	[-0.19, -0.74e- 1]	[0.31e- 2, 0.73e -2]

$$R^2=0.9938233721 \qquad F=281.5761182 \qquad p=0.8293753939e-7 \qquad f=0.7721667222e-1$$

由第三产业的时间序列曲线拟合及各参数表可以得出：第三产业  $x_7$  的回归方程的决定系数  $R^2$  在 0.99 以上，可靠度极高；且  $\beta$  置信区间均不包含零点，各变量对因变量影响显著； $F$  值均远远超过  $F$  检验的临界值； $P$  远小于  $\alpha$ 。

所以可得到：  $x_7=32.91-0.63t+0.92t^2 -0.13t^3 +0.0052t^4$  (式 4-4-4)

将  $x_4$ 、 $x_5$ 、 $x_7$  所得模型带入到建立的数学模型式 4-4-1 中，可以得到关于时间  $t$  的数学模型。

$$y=19887.36+10.29t+246.01t^2 -28.87t^3 +1.15t^4 \tag{式 4-4-5}$$

根据程序的设计，当时间为 1996 年时， $t=1$ ，故当时间为 2009 年时， $t=14$ ，2010 年上半年对应  $t=14.5$ 。带入公式 4-4-5，计算结果如表 4-4-3：

表 4-4-3 2009-2010 年多元逐步回归模型的预测值

年份	城镇就业人口预测值（万人）
2009	33300.61
2010 上半年	34686.73

也可通过系数值  $\beta$  确定各因素的时间序列拟合方程, 并做出对 2009 和 2010 上半年各因素值的预测, 结果如下:

表4-4-4 2009年和2010上半年各因素值预测结果

2010 年各相关因素值预测结果			
	国家财政收入 (亿元)	社会消费品零售总额 (亿元)	第三产业 (所占比值)
	$x_4$	$x_5$	$x_7$
2009 年	65053.35423	108887.1210	46.4568
2010 上半年	69836.88000	114938.1286	49.6446

根据程序的设计, 当时间为1996年时,  $x=1$ , 故当时间为2009年时,  $x=14$ . 2010年上半年的数据为14.5。

将各个数值代入式4-4-1, 可得2009年的就业人口数33300.79万人, 2010年上半年的就业人口数34686.94万人, 如表4-4-4所示:

表 4-4-5 2009-2010 年多元逐步回归模型的预测值

年份	城镇就业人口预测值 (万人)
2009	33300.79
2010 上半年	34686.94

两种方法得出预测结果很吻合, 证明建立的与时间相关的模型是正确的。

#### 4.4.3 BP神经网络模型

##### 1. 模型概述<sup>[13]</sup>

人工神经网络是由大量的神经元互联组成, 模拟大脑神经处理信息的方式并对信息进行并行处理和非线性转换的系统, 通过样本信息对神经网络的训练, 使其具有与大脑相类似的记忆、辨识能力, 按照网络的拓扑结构和运行方式, 神经网络可分为没有反馈的前向网络和相互结合性网络。前向网络由输入层、中间层 (隐含层) 和输出层组成, 中间层还可有若干层, 每层的神经元只接受前一层的输出。BP 网络是一种单向传播的多层前向网络, 其输入信号从输入层经隐含层逐层传播, 并传到输出层, 每一层神经元的状态只影响下一层神经元的状态。若输出层得不到期望输出, 则转为反向传播, 即误差信号沿着原来的连接通路返回, 并修改各层神经元的连接权值。两个传播过程反复交替使误差变得最小或小于某一期望值。目前, 在人工神经网络的实际应用中, 大部门采用 BP 网络及它的变化形式。下面是基本 BP 神经网络对我国城镇就业人口数的预测。

##### 2. 模型准备

主要因素的确定: 由上面回归模型的分析计算可知, 影响我国城镇就业人口的主要因素是财政收入、第三产业、全社会消费品零售总额。

##### 3. 模型的建立

用神经网络处理上述数据分两个阶段进行, 第1阶段是监督学习, 第2阶段

是无监督学习。监督学习室根据网络过去已知输入模式进行训练：将影响我国城镇就业人口数的三个元素作为输入单元，我国城镇就业人口数作为输出单元，采用 3—5—1 层结构，即输入层神经单元数为 3，输出层单元为 1，其中隐含层神经单元数目通过“试错法”，循环运算，确定隐含层神经单元数目为 5 效果较好，进行训练前对样本数据进行规范化处理，实输入和输出数据规范到[0, 1]区间，传递函数采用 sigmoid 型函数：

$$F(x)=1/(1+Exp(-\beta\times x)) \text{ , 常数 } \beta>0. \tag{式 4-4-6}$$

网络训练的误差目标函数  $E_k$ ，即

$$E_k = \frac{N}{2} \sum_{k=1}^N (u_k - v_k) \tag{式 4-4-7}$$

其中：N 为训练样本的数量；u 为网络的理想输出值；v 为网络的实际输出值。

利用 Levenberg-Marquardt 为训练网络，学习函数同样采用 Levenberg-Marquardt 训练数据，学习速率为 0.045，精度为 0.00015，选取 12 组数据中 1996-2004 年的 9 组实验数据作为训练数据，作为训练 10000 次达到精度标准，以剩余的三组数据作为预测数据。

预测结果如下：

表 4-4-6 BP 神经网络模型 2005—2007 年的预测结果

年份	城镇就业人口预测值（万人）	城镇就业人口准确值（万人）
2005	26890	27331
2006	27776	28310
2007	28725	29350

表 4-4-7 BP 神经网络模型对 2009 和 2010 年上半年的预测结果

年份	城镇就业人口预测值（万人）
2009	33018.98
2010 上半年	34332.15

通过计算预测结果与准确结果的误差率，分析得 2005-2007 年的城镇就业人口误差小于 5%。BP 模型不仅计算简便，便于计算机编程处理，程序的可移植性强，而且计算结果准确，能较地准确预测我国城镇就业人口数，但要强调的是由于网络的初始化是随机的，因此需要进行多次训练从而使精度符合要求。

#### 4.4.4 Logistic预测模型<sup>[14]</sup>



图 4-4-2 logistic 仿真模拟系统

比利时数学家 Verhulst 于 1838 年提出了著名的 Logistic 方程，最初用来预测种群在有限环境中的阻尼增长。自上世纪 60 年代以来，该方程被广泛用于动植物饲养栽培、医疗、人口、化学等众多领域。Logistic 方程可以表示为：

$$y = \frac{L}{1 + ae^{-bt}} \quad (\text{式 4-4-8})$$

式中：y 为待测生物的某生长指标；t 为生物生长的时间；L 为 y 增长上限，

$a = [\frac{L}{y_0} - 1]e^{bt_0}$  (其中  $y_0 = y(t)$ ，为 y 的初始值)；b 为待测生物的自由增长率。将公式 4-4-8 作变化可得

$$y = \frac{1}{\frac{1}{L} + \frac{a}{L}e^{-bt}} \quad (\text{式 4-4-9})$$

令  $b_0 = \frac{a}{L}$ ， $b_1 = e^{-b}$ ，可得

$$y = \frac{1}{\frac{1}{L} + b_0 b_1^t} \quad (\text{式 4-4-10})$$

城镇就业人口数是受社会需求和经济技术水平影响的动态系统，随着它们的提高，城镇就业人口数也在不断提高，但这种提高是有阈限的。城镇就业人口数总是受限于一定的社会需求和经济技术水平。在整个时间过程，呈阻尼增长。鉴于以上特点，可以近似用生态学中的 Logistic 曲线来预测城镇就业人口增长过程。

运用第二问中方法二主成分分析法求出的主成分 z1、z2 的历年值用 Logistics 曲线拟合，可以分别求得 z1 和 z2 的 logistics 方程，即他们与时间的关系：

$$z_1 = \frac{1}{\frac{1}{L_1} + b_{0_1} b_{1_1}^t} \quad (\text{式 4-4-11})$$



$$z_2 = \frac{1}{\frac{1}{L_2} + b_{0_2} b_{1_2}^t} \quad (\text{式 } 4-4-12)$$

将要预测的年份带到公式 4-4-11, 4-4-12, 即可分别得到  $z_1$  和  $z_2$  的值, 然后用第二问多元线性回归所得的城镇就业人口数与  $z_1$ 、 $z_2$  的函数关系, 求出该年城镇就业人口数, 达到预测的效果。

表 4-4-8 2009-2010 年 Logistic 模型的预测值

年份	城镇就业人口预测值 (万人)
2009	34198.87
2010 上半年	35286.51

这样的预测结果与前面几种模型相比, 数据差异不大, 比较吻合, 说明 Logistic 模型预测我国城镇就业人口问题效果还是较理想的。

## 5、提出建议

根据西方就业理论<sup>[15]</sup>, 失业分为摩擦性失业、结构性失业和周期性失业。摩擦性失业指劳动力供求在时空上的不一致引起的失业, 结构性失业指产业结构的变化导致的劳动力供求在结构上的不一致引起的失业, 周期性失业指经济的周期性变化至经济陷入衰退期时有效需求不足引起的失业。摩擦性失业与结构性失业构成自然失业, 充分就业指不存在周期性失业仅存在自然失业。摩擦性失业不可避免, 但可以通过培育劳动力市场等措施减少; 结构性失业也不可避免, 但可以通过劳动力培训等措施减少; 周期性失业可以避免, 即使周期性失业发生, 也可以通过需求管理等措施消除。只要政府实施正确的宏观管理政策, 就可以实现充分就业的目标。

我国就业压力的主要是由我国特殊国情决定的历史因素和制度因素引起的<sup>[16]</sup>: ①劳动力自然增长带来的就业压力; ②两大转变 (即经济体制从计划经济向社会主义市场经济转变, 经济增长方式从粗放型向集约型转变) 带来的就业压力; ③就业机制转变带来的就业压力; ④产业结构调整带来的就业压力: 我国长期实行重工业优先的产业发展战略。与发达国家及许多发展中国家相比, 第一产业比重偏高, 第二产业比重偏低。而从吸纳劳动力的能力看, 第一产业低于第二产业, 第二产业低于第二产业。下面结合前几题的研究成果和我国的具体国情, 在调整模型变量比较变动幅度的基础, 对增加我国城镇就业人数提出如下建议:

A. 从前面的分析得知, 国家财政收入是影响就业的主要因素之一, 而财政收入是我们在考虑政策因素时选取的主要指标, 其主要组成部分就是税收。于是针对这一点我们提出如下建议:

➤ 采取各种公共政策, 特别是通过税收政策<sup>[17]</sup>对市场进行干预, 来扩大就业、减少失业。由于在经济体系中, 如果存在着公共产品、外部性、规模收益递增、风险和不确定等情况, 市场机制就无法引导资源配置达到最优状态, 劳动力资源得不到合理配置。这时就需要政府来干预。政府参与市场经济运行, 弥补市场缺陷的主要手段就包括支出 (购买性支出和转移性支出)、税收以及公债。

B. 从前面的分析得知, GDP、社会消费品总额、第三产业增加值是影响就业的主要因素, 而 GDP、第三产业增加值是我们在考虑经济、结构因素时选取的指标。于是针对这一点我们提出如下建议: 实施促进经济增长的宏观经济政

- 策，扩大就业机会。
- 实行积极的产业政策、财政经济措施，调整经济结构，改善就业状况，提高就业容量。主要包括：
    - a. 调整经济结构，大力发展第三产业，使之成为增加就业的重要领域<sup>[18]</sup>；
    - b. 鼓励发展劳动密集型产业、服务业，扶持中小企业，鼓励、支持和引导非公有制经济发展，增强就业带动作用。
  - 实施积极的外贸政策：大力发展我国对外贸易和国际经济合作，多方式、多渠道创造就业岗位。通过扩大出口、吸引外资及劳务输出等办法促进就业水平提高是一些国家通常的做法。
  - 完善投资政策<sup>[19]</sup>，发挥重大投资项目对就业的带动作用。政府在安排重大投资和确定重大建设项目时，应当发挥投资和重大建设项目带动就业的作用，增加就业岗位。
  - 继续实施促进协调发展的区域政策，促进地区间就业的均衡增长。继续实施区域发展总体战略，加大对中西部地区等的支持力度，鼓励区域协作，统筹协调不同地区就业的均衡增长。
- C. 国家宏观调控就业率是我国构建和谐社会的重要手段，对此我们提出如下建议：
- 完善劳动力市场，加强对劳动力供给的调控。具体措施有：
    - a. 继续实施计划生育政策，从源头上控制劳动力供给；
    - b. 加强就业服务体系，完善劳动就业市场；
    - c. 明确职能定位，为促进劳动力市场发育创造良好的外部环境<sup>[20]</sup>。①是加快法制建设步伐，规范劳动力市场秩序；②是建立和完善统筹就业的政策体系；③是适应市场就业要求，大力开展就业公共服务。
  - 改革社会保障制度，完善社会保障体系，使得社会对失业率的承受能力不断提高。并改革失业保障法律制度，变被动失业救济为积极的就业激励。
  - 建立失业预警制度，建立失业警戒线，加强对就业形势的控制
  - 大力推进再就业工程
    - a. 建立再就业服务中心；
    - b. 建立解困和再就业专项资金，保障下岗职工的基本生活，使下岗职工在自谋职业期间有一定的生活保障。
  - 要扶助就业困难群体，促进社会安定。要措施有：
    - a. 通过开发或直接购买公益性就业岗位、对吸收就业困难人员的企业实行就业补贴、为困难群体提供免费转岗培训和各种就业服务等帮助其实现就业和再就业；
    - b. 实行政府扶持与市场化运作相结合的创业扶持政策，改善创业环境，开展“创业培训”，鼓励个人创业并以创业带动就业；
    - c. 通过政府援助下的劳动力退出政策创造更多的就业岗位。
- D. 个体因素中劳动者的素质也是影响就业的因素，考虑中国特殊的国情，我们提出以下建议：
- 加强高等职业教育和职业技术培训体系，提高失业人员再就业技能，改善劳动力供给质量。由于我国教育发展落后于经济发展，我国的劳动力素质普遍较低，不能适应产业结构的转换对高素质劳动力的需求，为此必须大力发展教育。首先是要发展各类高等教育，扩大大学生就学率，既可提高人力资源的素质，又可相应推迟劳动力就业年龄，缓解部分就业压力。其次是要大力

发展各类职业教育和职业技能培训，来大力提高劳动者的科技文化和职业技能素质，使更多的劳动者能适应产业结构变化和技术升级对劳动者素质的更高要求，减少结构性失业。

- 转变就业观念，大力发展非正规就业形式，推行灵活多样的就业制度，建立灵活多样的用工形式，在一些领域实行临时工、小时工、季节工和弹性工作制。

## 6、参考文献

- [1]张晓蓬，马丽靖，我国就业率因素的实证研究，现代商贸工业，20 卷第 5 期：29-30，2008
- [2]郑海燕，对于我国失业问题影响因素的研究，观察思考，P12-13，2003
- [3]柴林如，中国对外直接投资对国内就业影响分析，河北经贸大学学报，29 卷第 3 期：55—58，2008
- [4]牛一，改革开放以来经济增长与就业的模型分析，经济理论问题，9：15-16，2005
- [5]刘新卫，就业增长率与经济增长率关系的预测模型，科技创新，12 期：168，2008
- [6]徐旭川，中国劳动生产率的就业效应分析，当代财经，2008，10：17-22，2008
- [7]金丹，对汇率与失业率关系的相关分析，石家庄法商职业教学与研究，4 卷 1 期：24-27，2008
- [8]程永宏，服务业就业比重与失业率的关系研究，中国软科学，2005 年 2 期：46—56，2005
- [9]范通达，江兵，基于主成分分析的多元线性回归用水量模型研究—以安徽省为例，基建优化，28 卷 2 期：53-55，2007
- [10]姜启源，谢金星，数学建模案例选集[M]，北京：高等教育出版社，2006
- [11]赵东方，数学模型与计算[M]，北京：科技出版社，2007
- [12]姜启源，谢金星，叶俊，数学建模[M]，北京：高等教育出版社，2005
- [13]全国大学生数学建模竞赛组委会编，数学建模的实践[C]，北京：高等教育出版社，2006
- [14]浦小松，基于 Logistic 回归模型的加拿大妇女就业影响因素的实证研究及对我国的启示，国际经济，9：27-31，2008
- [15]黄万阳，我国高失业率原因分析及对策建议，财经问题研究，4 期，2002
- [16]刘俊美，城镇失业问题及政府调控对策研究，大连理工大学[D]，2005
- [17]付伯颖，促进就业的财政政策研究，东北财经大学[D]，2007
- [18]杨宜勇，当前我国调控就业形势的主要措施选择，理论视野，2000. 2:28-31, 2000
- [19]王策，郭玉华，缓解我国就业压力的政策思考，地质技术经济管理，21 卷第 3 期：9-12，1999
- [20]郭蚌珠，曾毅红，明确政府职能，缓解就业压力，前线，19-20，2005
- [21] Allin Cottrell, David Begg's Full-employment Model, Economics 211
- [22] Elisabeta JABA , Carmen PINTILESCU2, A Forecast Model of the Employment Rate in Romania

## 7、附录

相关程序：

1、主要影响因素的散点图和拟合曲线

```
clear;
a=[19922    71176.59165    22913.55    24133.8    7407.99    28360.2
    19.5    33    7937.55    1704.25    105.049
20781    78973.035    24941.11    26967.2    8651.14    31252.9    18.1
    34.4    9233.56    1903.59    139.89
21616    84402.27977    28406.3    26849.7    9875.95    33378.1    17.3
    36.5    10798.18    2154.38    144.959
22412    89677.05475    29855.76    29896.2    11444.08    35647.9    16.2
    38    13187.67    2408.06    154.675
23151    99214.55431    32917.73    42183.6    13395.23    39105.7    14.8
    39.3    15886.50    2736.88    165.574
23940    109655.1706    37213.49    42183.6    16386.04    43055.4    14.1
    40.7    18902.58    3361.02    212.165
24780    120332.6893    43499.91    51378.2    18903.64    48135.9    13.5
    41.7    22053.15    3979.08    286.407
25639    135822.7561    55566.61    70483.5    21715.25    52516.3    12.5
    41.5    24649.95    4505.51    403.251
26476    159878.3379    70477.4    95539.1    26396.47    59501.0    13.1
    40.7    28486.89    5143.65    699.32
27331    159878.3379    88773.6    116921.8    31649.29    67176.6    12.5
    40    33930.28    6104.18    818.872
28310    211923.5    109998.2    140971.4    38760.2    76410.0    11.3    40
    40422.73    7425.98    1066.3
29350    249529.9    137323.9    166740.2    51321.78    89210.0    11.1
    40.4    49781.35    11793.96    1527.9
];
y=a(:,1);
x1=a(:,2);
x2=a(:,3);
x3=a(:,4);
x4=a(:,5);
x5=a(:,6);
```

```

x6=a(:, 7);
x7=a(:, 8);
x8=a(:, 9);
x9=a(:, 10);
x10=a(:, 11);

plot(x4, y, ' * ');
hold on;
wi=5000:50:60000;
p=polyfit(x4, y, 2);           %调用多项式拟和函数, 返回拟合参数
yc=polyval(p, wi);            %计算拟合值
plot(wi, yc, ' - ')           %画出计算值与拟合值的图形
title(' 就业人口数—国家财政收入 散点图');
xlabel(' 国家财政收入 (亿元) ');
ylabel(' 就业人口数 (万人) ');

```

```

plot(x5, y, ' * ');
hold on;
wi=28000:50:90000;
p=polyfit(x5, y, 2);           %调用多项式拟和函数, 返回拟合参数
yc=polyval(p, wi);            %计算拟合值
plot(wi, yc, ' - ')           %画出计算值与拟合值的图形
title(' 就业人口数—全社会消费品零售总额 散点图');
xlabel(' 全社会消费品零售总额 (亿元) ');
ylabel(' 就业人口数 (万人) ');

```

## 2、多元回归模型的求解

```

a=[1 1 1 1 1 1 1 1 1 1 1 1]';
x=[a x4, x5, x7];
[b, bint, r, rint, stats]=regress(y, x, 0.05)
vpa(b, 10)
y2=b(1)+b(2)*x4+b(3)*x5+b(4)*x7;
plot(1996:2007, y, ' * ', 1996:2007, y2(:));
xlabel(' 年份 ');
ylabel(' 就业人数/万人 ');
%残差分析
figure;
plot(1996:2007, y-y2);
xlabel(' /年份 ');
ylabel(' /万人 ');
title(' 就业人数拟合残差图 ');

```

## 3、西部地区城镇就业人数拟合图

```
clear
```

```

a=[3462 6984.8 7158.8 6169.9 101.5
3462 7760.4 8515.4 6675.2 99.8
3628 8668.2 10843.5 7205 101.4
3710 9986.4 13754.4 8031 103.6
3896 13237.1 17645 8783 101.4
4036 15251.4 21996.9 9728 101.7
4340.4 9850.3 28250.9 11309 105.4
];
y=a(:,1);
x1=a(:,2);
x2=a(:,3);
x3=a(:,4);
x4=a(:,5);
l=[1 1 1 1 1 1 1]';
x=[1 x1,x2,x3,x4,x1.^2,x2.^2,x3.^2,x4.^2];
[b,bint,r,rint,stats]=regress(y,x,0.05);
y2=b(1)+b(2)*x1+b(3)*x2+b(4)*x3+b(5)*x4+b(6)*x1.^2+b(7)*x2.^2+b(8)*x3.^2+b(9)*x4.^2;
plot(2001:2007,y,'*r',2001:2007,y2,'b',2001:2007,y2,'.')
xlabel('年份');
ylabel('就业人数/万人');
title('西部地区城镇就业人数拟合图');
vpa(b,15)
vpa(bint,10)

```

#### 4、第三产业指标的预测模型

```

clear;
a=[19922 71176.59165 22913.55 24133.8 7407.99 28360.2
19.5 33 7937.55 1704.25 105.049
20781 78973.035 24941.11 26967.2 8651.14 31252.9 18.1
34.4 9233.56 1903.59 139.89
21616 84402.27977 28406.3 26849.7 9875.95 33378.1 17.3
36.5 10798.18 2154.38 144.959
22412 89677.05475 29855.76 29896.2 11444.08 35647.9 16.2
38 13187.67 2408.06 154.675
23151 99214.55431 32917.73 42183.6 13395.23 39105.7 14.8
39.3 15886.50 2736.88 165.574
23940 109655.1706 37213.49 42183.6 16386.04 43055.4 14.1
40.7 18902.58 3361.02 212.165
24780 120332.6893 43499.91 51378.2 18903.64 48135.9 13.5
41.7 22053.15 3979.08 286.407
25639 135822.7561 55566.61 70483.5 21715.25 52516.3 12.5
41.5 24649.95 4505.51 403.251
26476 159878.3379 70477.4 95539.1 26396.47 59501.0 13.1

```

```

    40.7    28486.89    5143.65    699.32
27331 159878.3379    88773.6    116921.8    31649.29    67176.6    12.5
    40    33930.28    6104.18    818.872
28310 211923.5    109998.2    140971.4    38760.2    76410.0    11.3    40
    40422.73    7425.98    1066.3
29350 249529.9    137323.9    166740.2    51321.78    89210.0    11.1
    40.4    49781.35    11793.96    1527.9
];
% plot(1:12, a(:, 8), '*' )

xx=[1:12]';
yy=a(:, 8);
l=[1 1 1 1 1 1 1 1 1 1 1 1]';
xx1=[1, xx, xx.^2, xx.^3, xx.^4];
[b, bint, r, rint, stats]=regress(yy, xx1, 0.05)
vpa(b, 10)
y2=b(1)+b(2)*xx+b(3)*xx.^2+b(4)*xx.^3+b(5)*xx.^4;
%对 x1 预测函数的图示
xxx=[1:0.01:13]';
yyy=b(1)+b(2)*xxx+b(3)*xxx.^2+b(4)*xxx.^3+b(5)*xxx.^4;
plot(1:0.01:13, yyy)
hold on
plot(1:12, yy, '*', 1:12, y2, 'o')
xlabel(' 年/(1996 年对应第一年)');
ylabel(' 第三产业比例/%');
vpa(stats, 10)

```