第一大题：解：

作出描述碳含量-合金强度之间关系的数据散点图和折线图：

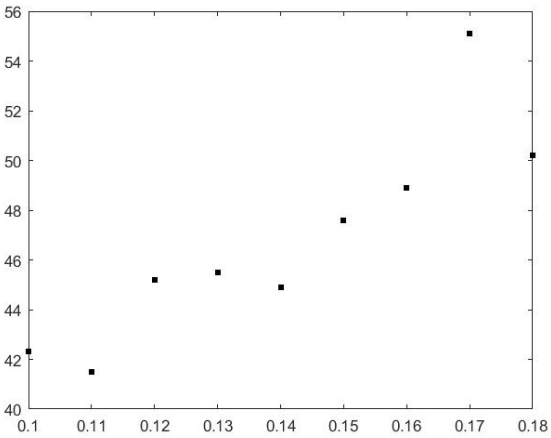


图 42-0703_1-1 碳含量-合金强度数据组散点图

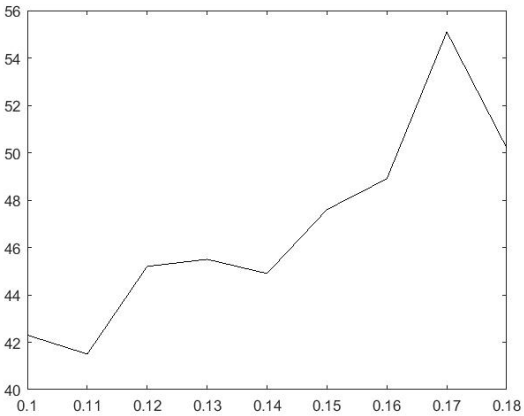


图 42-0703_1-2 碳含量-合金强度数据组折线图

显然地，我们指出：数据点（0.17，55.1）显著地偏离了数据点的整体趋势，认为此数据点存在缺陷，应予舍去；同时指出，此模型可以用一次或二次多项式（即 $y=ax+b$ 或 $y=ax^2+bx+c$）进行拟合。

我们先从一次多项式的情形开始计算：

依据最小二乘原则，有：

$$\begin{cases} a=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n}x_i^2} \\ b=\bar{y}-a\bar{x} \end{cases} \quad (1)$$

其中：

$$\begin{cases} \bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n} \\ \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} \end{cases} \quad (2)$$

引用 MATLAB 命令进行计算，得：

$$\begin{cases} y = 108.05x + 31.04 \\ R^2 = 0.9957 \end{cases} \quad (3)$$

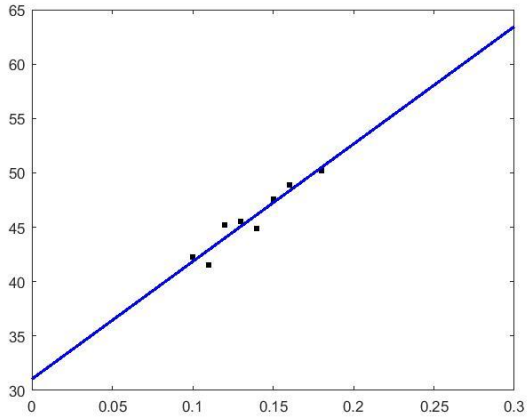


图 42-0703_1-3 碳含量-合金强度关系的线性预测

根据拟合系数判断，此模型拟合效果甚佳。

下面我们对二次多项式的情形进行讨论：

引用 MATLAB 命令直接进行多项式拟合得：

$$\begin{cases} y = -79.13x^2 + 130.08x + 29.56 \\ R^2 = 0.9978 \end{cases} \quad (4)$$

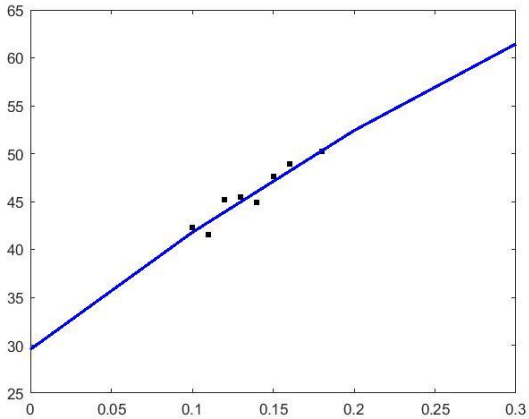


图 42-0703_1-3 碳含量-合金强度关系的二次多项式预测

由拟合系数，我们指出，二次多项式进行拟合的结论优于一次多项式拟合的结论。事实

上，二者的精确程度相差并不明显。我们也不能明确地指出这种精确来源于参数的增多抑或是数据本身的趋势。

证完。

第三大题：解：

规范：本题中以 $y$ 指代人口数，单位为亿人；以 $x$ 指代时间，单位为年。

我们优先给出人口-时间数据的散点图和折线图：

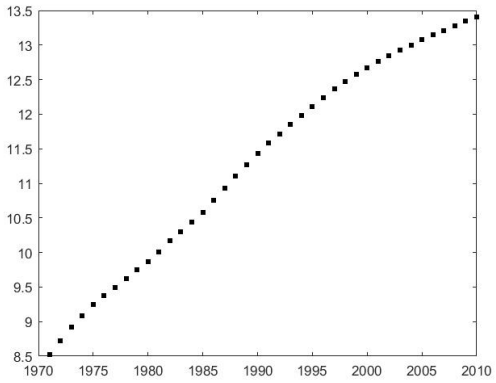


图 42-0701_3-1 人口-时间数据组散点图

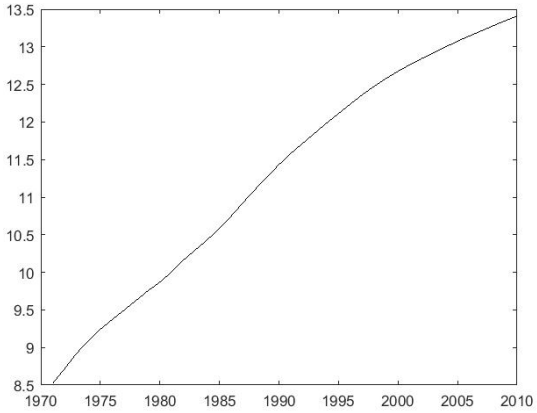


图 42-0701_3-2 人口-时间数据组折线图

国家统计局给出的 2011-2018 年间中国人口总数为：

| 年份 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|---|---|---|---|---|---|---|---|---|
| 人口总数/亿 | 13.4735 | 13.5405 | 13.6072 | 13.6782 | 13.7462 | 13.8271 | 13.9008 | 13.9538 |

2019 年全国人口总量为 14.0005 亿人。

描述人口状态的常用预测模型一般是 *Malthus* 模型和 *Logsitic* 模型，但是也不排除用一次或者二次函数进行拟合的可能性。我们将分别讨论这四种预测模型的优劣程度。

（1） *Malthus* 模型：

即使用函数 $y = Ae^{bx}$（或者使用另一种描述形式， $y = e^{bx+a}$ ）进行描述。

这种情形下我们一般采用变量替换 $u = \ln y$ 将非线性函数转化为线性函数，然后依照最小二乘原则进行求解：

$$\begin{cases} a = \dfrac{\sum_{i=1}^{n} x_i u_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2} \\ b = \bar{u} - a\bar{x} \end{cases} \quad (1)$$

其中：

$$\begin{cases} \bar{u} = \dfrac{\sum_{i=1}^{n} \ln y_i}{n} \\ \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} \end{cases} \quad (2)$$

引用 MATLAB 命令进行求解，得：

$$u = 0.0117x - 20.8907 \quad (3)$$

即：

$$\begin{cases} u = e^{0.0117x - 20.8907} \\ R^2 = 0.9707 \end{cases} \quad (4)$$

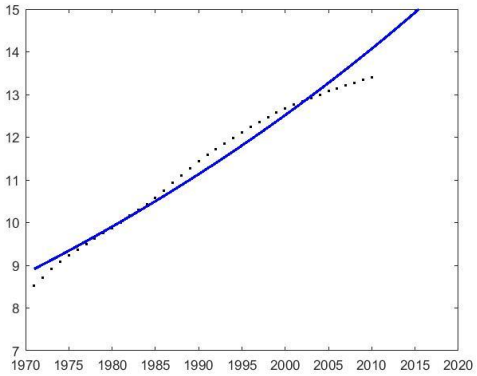


图 42-0701_3-3 人口-时间关系的 *Malthus* 预测模型图示

我们将后十年的预测数据与实际数据作对比，并计算残差平方和，得：

$$\sum_{i=1}^{9} \Delta y^2 = 13.2508 \quad (5)$$

同时事实上，根据图像可以显而易见地看出，2005 年后 *Malthus* 预测模型的演进趋势已经与实际情形产生了显著的差异，因此此处并不建议使用 *Malthus* 预测模型。

（2）*Logsitic* 模型：

*Logsitic* 模型是基于微分方程：

$$\frac{dy}{dx} = ry\left(1 - \frac{y}{k}\right) \quad (6)$$

进行的回归分析。这一微分方程的通解为：

$$y=\frac{KY_0e^{rx}}{K+Y_0(e^{rx}-1)} \quad (7)$$

其中 $K$ 指环境容纳量，$r$ 指无约束情形下的自然增长率，$Y_0$ 指人口总量初值。在进行回归分析时我们一般确定参量 $K$，然后利用最小二乘原则对 $r$，$Y_0$ 进行计算。

考虑到在时间原点（$x=0$，即公元元年）处的中国人口初值远小于 1 亿，在建立模型时我们对模型作出适当的简化，即忽略掉函数分母位置的 $Y_0$ 因子。这样，通解则被简化为：

$$y=\frac{KY_0e^{rx}}{K+Y_0e^{rx}} \quad (8)$$

在进行回归分析时，指出：本题中 $Logsitic$ 模型的环境容纳量指标，即参量 $K$ 的大小，是基于 1991 年中国科学院自然资源综合考察委员会的《中国土地资源生产能力及人口承载力研究》中的表述“我国人口承载量最高应控制在 16 亿左右”。即：

$$K=16 \quad (9)$$

作变量替换：

$$Y=\ln(\frac{1}{y}-\frac{1}{K}) \quad (10)$$

以最小二乘法为基础计算参量 $r$ 和 $Y_0$：

$$\begin{cases} r=0.0402 \\ Y_0=e^{-76.2477} \end{cases} \quad (11)$$

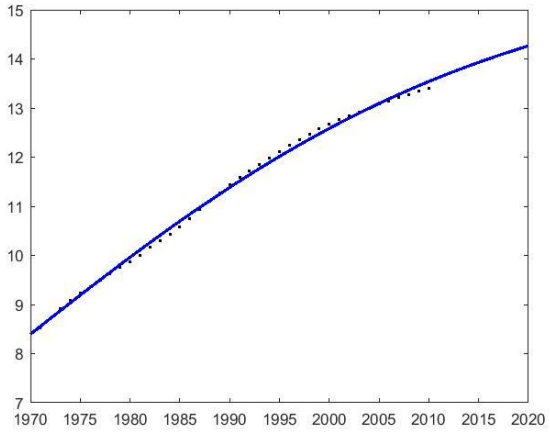


图 42-0701_3-4 人口-时间关系的 $Logistic$ 预测模型图示

这一模型的回归系数为：

$$R^2=0.9974 \quad (12)$$

后十年的预测数据与实际数据之间的残差平方和为：

$$\sum_{i=1}^{9}\Delta y^2 = 0.2778 \quad (13)$$

事实证明经过简化的 *Logsitic* 模型在本题中可以相对很好地描述和预测这 50 年的人口走向，是本题中最符合实际、最接近真实情形的数学模型。然而这一模型仍存在缺陷：这一模型最初是用于描述动物种群的种群密度的演化趋势的，只考虑了自然条件约束而忽略了其他的隐性约束。而在人类社会中，这种隐性的因素——包括社会生活和社会习俗，国家指令性政策和计划等——并没有被这一方程组直观地体现。当然，这些因素的数学描述一向是较为困难的，因此我们不必过于求全责备而对这一模型进行各种纷杂繁复的修正。

（3）线性模型：

我们将直接采用最小二乘法计算线性模型的系数：

$$\begin{cases} a = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2} \\ b = \bar{y} - a\bar{x} \end{cases} \quad (14)$$

其中：

$$\begin{cases} \bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n} \\ \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} \end{cases} \quad (15)$$

引用 MATLAB 进行计算，得：

$$\begin{cases} y = 0.1295x - 246.4248 \\ R^2 = 0.9872 \end{cases} \quad (6)$$

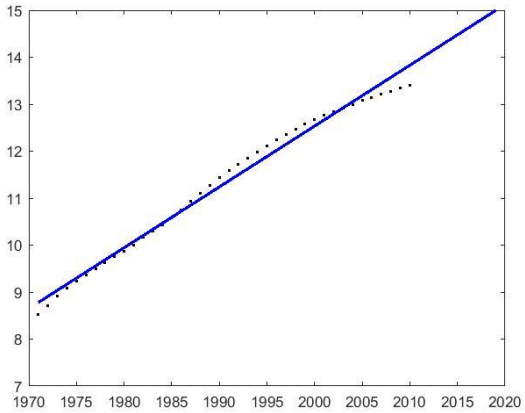


图 42-0701_3-5 人口-时间关系的线性预测模型图示

将后十年的预测数据与实际数据作对比，并计算残差平方和，得：

$$\sum_{i=1}^{9}\Delta y^2 = 5.0098 \quad (17)$$

其残差和显著地小于 *Malthus* 预测模型。然而线性预测模型存在显著的缺陷：一方面，它只能线性地近似描述较短的一个历史周期（大约 30-70 年）的人口变化趋势，对于更长的历史周期（200 年及以上）则会产生显著的误差，甚至是有悖于常理的结论（如，负数人口）。另外，完全线性的增长也不符合人口-时间演进规律的常理和常识。因此，虽然这一模型在本题中优于 *Malthus* 预测模型，但是其科学性和准确性却是本题四个模型中最低的，也是最不建议使用的一种数学模型。

（4）二次函数模型：

采用变量替换：$x' = x - 1970$ （18）

直接引用 MATLAB 中的回归函数进行回归分析，并将原变量回代得：

$$\begin{cases} y = -0.0013(x-1970)^2 + 0.1812(x-1970) + 8.2882 \\ R^2 = 0.9971 \end{cases} \quad (19)$$

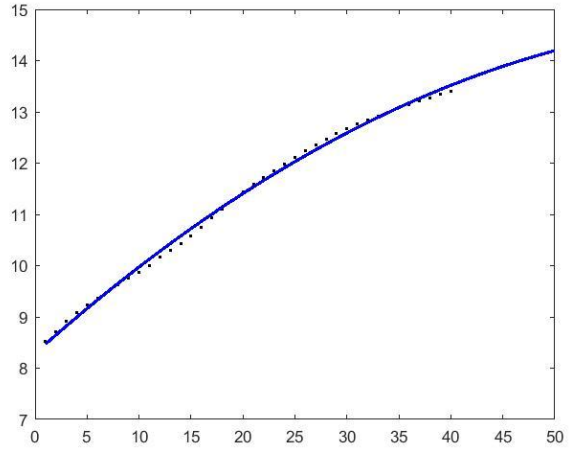


图 42-0701_3-6 人口-时间关系的二次多项式预测模型图示

计算得到的后十年的预测数据与实际数据之间的残差平方和为：

$$\sum_{i=1}^{9}\Delta y^2 = 0.1575 \quad (20)$$

虽然此模型残差和亦显著小于 *Malthus* 预测模型。这一模型仍存在与线性预测模型类似的缺陷：一方面，这种二次多项式模型也只能对人口-时间的演进关系作短期（大约 30-70 年）的简单预测，而对长期预测无能为力，甚至会产生不合逻辑的结论（负人口数）；另一方面，二次多项式模型中各参量的统计意义不明确，无法对这种人口变化的社会现象作出科学的、合理的解释。因此，这种二次多项式模型不能作为一个普适的、长期的预测，而对于短期的人口变化反而可以作出一个相当接近的估计。这一论断在之前对于美国人口的预测中也得到了充分的体现——在 *Malthus* 模型、*Logsitic* 模型、线性模型、二次多项式模型中，二次多项式模型在预测后十年美国人口总数时得到了最小的残差平方和。

*注意：在引用三次多项式预测模型进行预测时：

采用变量替换：$x'=x-1970$（13）

直接引用 MATLAB 中的回归函数进行回归分析，并将原变量回代得：

$$\begin{cases} y=-0.0001(x-1970)^3+0.0023(x-1970)^2+0.1223(x-1970)+8.5015 \\ R^2=0.9993 \end{cases} \quad (21)$$

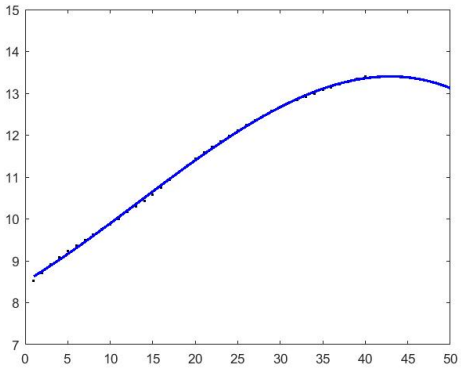


图 42-0701_3-7 人口-时间关系的三次多项式预测模型图示

计算得到的预测-实测残差方和为：

$$\sum_{i=1}^{9}\Delta y^2=1.9792 \quad (22)$$

在引用更高次的多项式进行回归分析时观测到，拟合系数 $R^2$ 随多项式次数的提高而逐渐缩小，预测-实测残差方和反而在随着多项式次数的增大而增大。如：四次多项式的预测-实测残差方和已经达到了：

$$\sum_{i=1}^{9}\Delta y^2=2.1304 \quad (23)$$

指出：多项式次数的增大使得模型在指定的拟合区间内的精确度逐渐增大，其原理类似于降秩线性方程组的求解——方程组秩与变量数的差值越小，解空间元素的分布越集中，可能的解的情形越少。亦即，拟合时多项式次数的增大使得这种回归分析逐渐向 *Lagrange* 插值法逼近。而这种类似于“插值”逼近只能提高模型在指定区间的精度而不能提高模型的预测精度——次数的提高甚至会有害于这种预测。这种现象出现的本质，在于多项式中各参数的物理意义的不确定性。同时考虑到 *Lagrange* 插值法中 *Runge* 现象的存在，我们并不建议使用三次及以上的多项式对于实际问题进行拟合。

**根据前文的分析，作者定性地指出：二次多项式模型是进行短期人口预测（10-15 年）的最精确的多项式函数预测模型。

证完。

第五大题：解：

作出距离-矿物丰度之间的数据散点图和折线图：

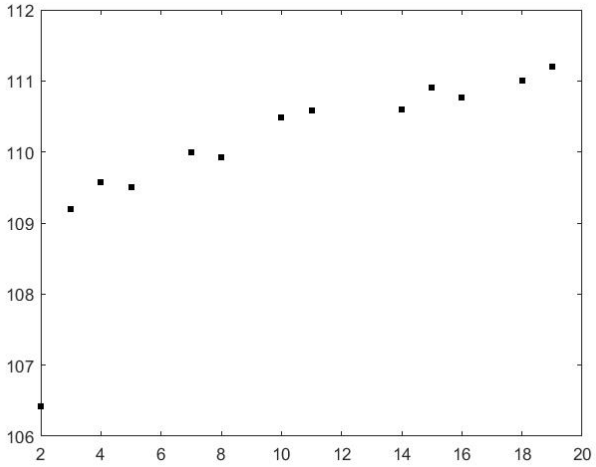


图 42-0703_5-1 距离-矿物丰度数据组散点图

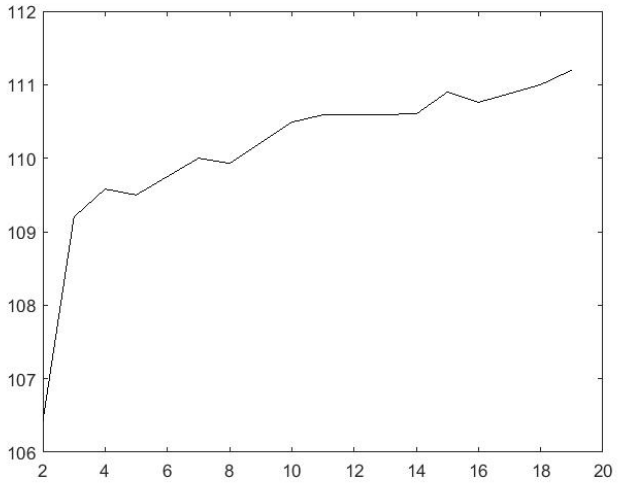


图 42-0703_5-2 距离-矿物丰度数据组散点图

显然地，第一个数据（2，106.42）显著地偏离的总体趋势，因此我们推断此数据存在问题，应予舍去；同时推断，此模型适于使用线性函数 $y=ax+b$ 拟合预测。

采用最小二乘法计算：

$$\begin{cases} a=\dfrac{\sum\limits_{i=1}^{n}x_i y_i-n\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{n}x_i^2} \\ b=\overline{y}-a\overline{x} \end{cases} \quad (1)$$

其中：

$$\begin{cases} \overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n} \\ \overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} \end{cases} \quad (2)$$

引用 MATLAB 进行计算，得：

$$\begin{cases} y = 0.144x + 109.0778 \\ R^2 = 0.9976 \end{cases} \quad (5)$$

根据回归系数 $R^2$ 判断，该模型拟合效果甚佳。

作图，得：

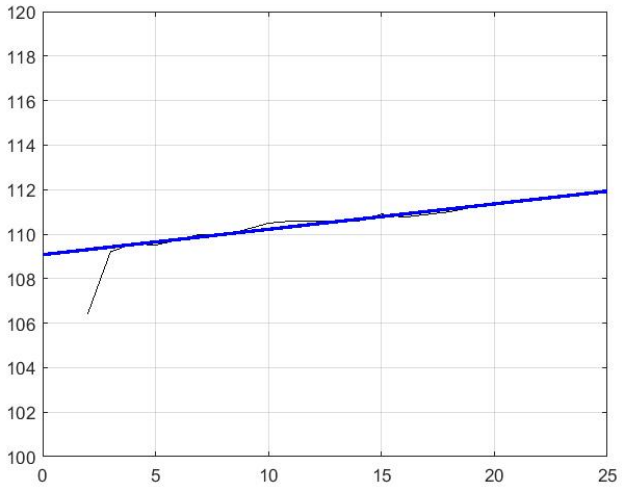


图 42-0703_5-3 距离-矿物丰度预测图

证完。

附录：

（一）第一大题的 MATLAB 回归分析代码：

```
clear;clc
x=[0.1:0.01:0.16,0.18];
y=[42.3,41.5,45.2,45.5,44.9,47.6,48.9,50.2];
axis([0.1 0.3 40 60])
grid
plot(x,y,'k.','markersize',10);
hold on
n1=polyfit(x,y,2);
x0=0:0.1:0.3;
plot(x0,polyval(n1,x0),'b-','linewidth',2);
yp=0;
for i=1:8
    yp=yp+y(i)/8;
end
y1=0;
yd=0;
for j=1:8
    y1=y1+(y(i)-yp).^2;
    yd=yd+(y(i)-polyval(n1,x(i)))^2;
end
r=1-yd/y1
```

（二）第三大题中对 *Malthus* 模型和 *Logsitic* 模型进行回归分析的 MATLAB 代码：（此处引用的代码为分析 *Logsitic* 模型时所用的代码，稍作调整即可用于 *Malthus* 模型。）

```
clear;clc;
t=1971:2010;
tex=2011:1:2019;
m=[8.5229,8.7177,8.9221,9.0859,9.2420,9.3717,9.4974,9
.6259,9.7542,9.8705,10.0072,10.1654,10.3008,10.4357,1
0.5851,10.7507,10.9300,11.1026,11.2704,11.4333,11.582
3,11.7171,11.8517,11.985,12.1121,12.2389,12.3626,12.4
761,12.5786,12.6743,12.7627,12.8453,12.9227,12.9988,1
3.0756,13.1448,13.2129,13.2802,13.3474,13.41];
mex=[13.4735,13.5405,13.6072,13.6782,13.7462,13.8271,
13.9008,13.9538,14.0005];
plot(t,m,'k.','markersize',5);
hold on
axis([1970 2020 7 15])
nm=polyfit(t,log(1./m-1/16),1);
t0=1970:0.1:2020;
```

```
plot(t0,1./(exp(polyval(nm,t0))+1/16),'b-
','linewidth',2)
ma=0;
for i=1:40
    ma=ma+m(i)/40;
end
md=0;
mdm=0;
for i=1:40
    md=md+(m(i)-ma).^2;
    mdm=mdm+(m(i)-
1./(exp(polyval(nm,t(i)))+1/16)).^2;
end
r=1-mdm/md;
delex=0;
for i=1:9
    delex=delex+(mex(i)-
1./(exp(polyval(nm,tex(i)))+1/16)).^2;
end
```

（三）第三大题中多项式回归分析的 MATLAB 代码:（此处以三阶多项式为示例，调整参数之后可以用于拟合其他阶数的多项式）

```
clear;clc;
t=1:40;
tex=41:1:49;
m=[8.5229,8.7177,8.9221,9.0859,9.2420,9.3717,9.4974,9
.6259,9.7542,9.8705,10.0072,10.1654,10.3008,10.4357,1
0.5851,10.7507,10.9300,11.1026,11.2704,11.4333,11.582
3,11.7171,11.8517,11.985,12.1121,12.2389,12.3626,12.4
761,12.5786,12.6743,12.7627,12.8453,12.9227,12.9988,1
3.0756,13.1448,13.2129,13.2802,13.3474,13.41];
mex=[13.4735,13.5405,13.6072,13.6782,13.7462,13.8271,
13.9008,13.9538,14.0005];
plot(t,m,'k.','markersize',5);
hold on
axis([0 50 7 15])
nm=polyfit(t,m,3)
t0=0:0.1:50;
plot(t0,polyval(nm,t0),'b-','linewidth',2)
ma=0;
for i=1:40
    ma=ma+m(i)/40;
end
md=0;
```

```
mdm=0;
for i=1:40
    md=md+(m(i)-ma).^2;
    mdm=mdm+(m(i)-polyval(nm,t(i))).^2;
end
r=1-mdm/md
delex=0;
```

（四）第五大题的 MATLAB 线性回归分析代码：

```
clear;clc
x=[3,4,5,7,8,10,11,14,15,16,18,19];
y=[109.2,109.58,109.5,110.0,109.93,110.49,110.59,110.
60,110.9,110.76,111,111.2];
axis([0 25 100 120])
grid
plot(x,y,'k-','markersize',10);
n1=polyfit(x,y,1);
x0=0:0.1:25;
plot(x0,polyval(n1,x0),'b-','linewidth',2);
hold on
yp=0;
for i=1:12
    yp=yp+y(i)/12;
end
y1=0;
yd=0;
for j=1:12
    y1=y1+(y(i)-yp).^2;
    yd=yd+(y(i)-polyval(n1,x(i)))^2;
end
r=1-yd/y1
```

本文作者：机 916 班 杨逢诜；机 928 班 林晓龙；机 936 班 李昊天；机 937 班 赵明智.