目录

回归模型

1 基础知识

回归是研究自变量（解释变量）和因变量（被解释变量）之间的函数关系，按照自变量的个数和回归函数的类型可分为一元线性回归，一元非线性回归，多元线性回归，多元非线性回归；还有在回归过程中可以调整自变量数量的回归方法叫逐步回归；输出值只有两个的对数几率回归。

2 多元线性回归版本 1

财政收入,是指政府部门为完成其职责、实现公共政策目标以及供给公共物品和公共服务需要,所筹集的所有资本的总额。对一九九零年至二零一三年的统计资料加以分析(本项目所用数据均来自《统计年鉴》),统计结果如表 1:

表 1 1990-2013 年国家财政收入数据

年份	工业总产值	农业总产值	建筑业总产值	社会商品零售总额	全国人口总数	国家财政收入
1990	6858	4954.3	859.4	8300.1	114333	2937.1
1991	8087.1	5146.4	1015.1	9415.6	115823	3149.48
1992	10284.5	5588	1415	10993.7	117171	3483.37
1993	14188	6605.1	2266.5	14270.4	118517	4348.95
1994	19480.7	9169.2	2964.7	18622.9	119850	5218.1
1995	24950.6	11884.6	3728.8	23613.8	121121	6242.2
1996	29447.6	13539.8	4387.4	28360.2	122389	7407.99
1997	32921.4	13852.5	4621.6	31252.9	123626	8651.14
1998	34018.4	14241.9	4985.8	33378.1	124761	9875.95
1999	35861.5	14106.2	5172.1	35647.9	125786	11444.08
2000	40033.6	13873.6	5522.3	39105.7	126743	13395.23
2001	43580.6	14462.8	5931.7	43055.4	127627	16386.04
2002	47431.3	14931.5	6465.5	48135.9	128453	18903.64
2003	54945.5	14870.1	7490.8	52516.3	129227	21715.25
2004	65210	18138.4	8694.3	59501	129988	26396.47
2005	77230.8	19613.4	10367.3	67176.6	130756	31649.29
2006	91310.9	21522.3	12408.6	76410	131448	38760.2
2007	110534.9	24658.1	15296.5	89210	132129	51321.78
2008	130260.2	28044.2	18743.2	114830.1	132802	61330.35
2009	135239.9	30777.5	22398.8	132678.4	133450	68518.3
2010	160722.2	36941.1	26661	156998.4	134091	83101.51
2011	188470.2	41988.6	31942.7	183918.6	134735	103874.43
2012	199670.7	46940.5	35491.3	210307	135404	117253.52
2013	210689.4	51497.4	38995	237809.9	136072	129209.64

根据表 1 的我国财政收入数据构建多元线性回归模型。

2.1 相关性分析

假设我国的财政收入为y，各影响性指数工业总产值，农村地区生产总值,我国建筑业总量,中国社会产品零售总额,全国人口总数分别为$x_i, i=1,2,...,5$。要大致地研究y与x_i的因果关系,必须首先开展相关性分析研究,目前主要采用的包括 Pearson 最大相关系数分析方法、Spearman 最小关联系数和 Kendall 最大相关系数分析方法。

Pearson 相关系数(Pearson Correlation Coefficient)是一种判断的二组统计整体相互之间是不是连在同一根线上,也用来判断二定距变量相互之间的直线联系。当二变量均为连续变量,且二变量的总体为正态分布以及二变量之间为线性关系时可使用 Pearson 的方法。二变量间的皮尔逊相关系数可定义为二变量间的协方差与标准差的比商：

假设随机变量x, y的样本方差为：

$$S_x^{\ 2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_y^{\ 2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

样本协方差为：

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

则定义样本相关系数为：

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_x}\sqrt{S_y}}$$

Kendall(肯德尔)系数的定义为：n个同类的统计对象按特定属性排序，其他属性通常是乱序的。同序对（concordant pairs）和异序对（discordant pairs）之差与总对数$n(n-1)/2$的比值定义为 Kendall(肯德尔)系数。

斯皮尔曼相关系数被定义成等级变量之间的皮尔逊相关系数，当两变量都是连续数据且两变量总体不为正态分布时可以采用斯皮尔曼分析，对于样本容量为n的样本，n个原始数据被转换成等级数据，相关系数ρ的公式表示为：

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

将数据标准化后利用 Matlab 绘制出散点图，

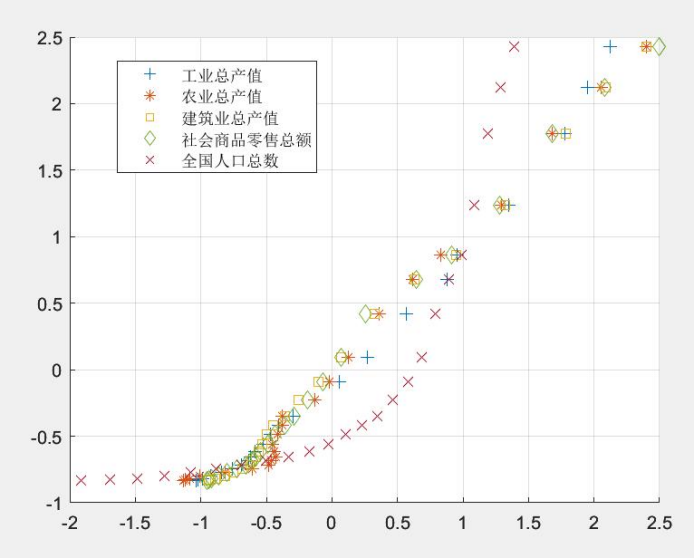


图 1 国家财政收入与各指标散点图

并计算出相关系数为：

表 2 相关系数结果

	x_1	x_2	x_3	x_4	x_5	y
x_1	1.00	0.99	0.99	0.99	0.88	0.99
x_2	0.99	1.00	0.99	0.99	0.87	0.99
x_3	0.99	0.99	1.00	1.00	0.83	1.00
x_4	0.99	0.99	1.00	1.00	0.84	1.00
x_5	0.88	0.87	0.83	0.84	1.00	0.82
y	0.99	0.99	1.00	1.00	0.82	1.00

并绘制出相关系数热力图：

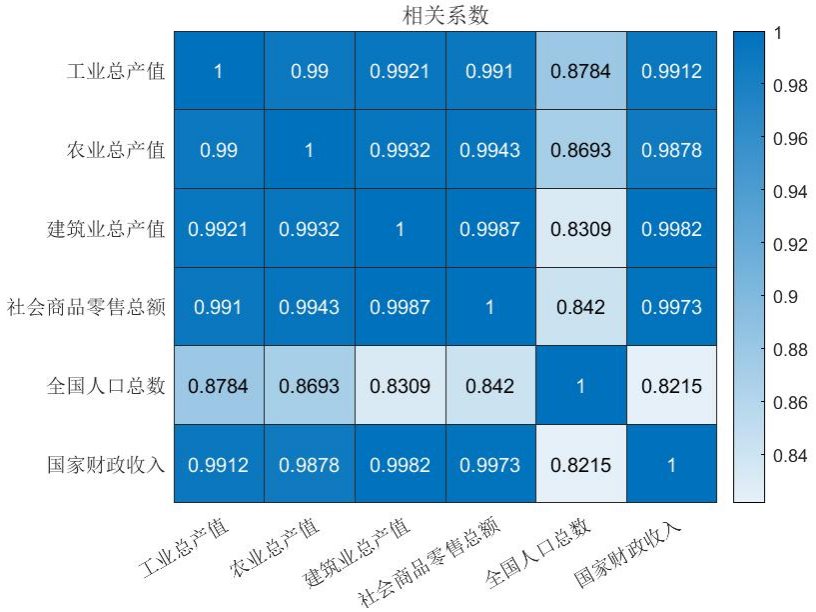


图 2 相关系数热力图

无论是从散点图还是相关系数表还是相关系数热力图都可以看出来，在我国财政收支和各指数相互之间具有着很大的线性关系。

2.2 模型的建立与求解

于是, 可以形成这样的多元线性回归模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\varepsilon$$

其中，x_1,x_2,x_3,x_4,x_5 称为回归变量，表示工业总产值，农业总产值，建筑业总产值，社会商品零售总额，全国人口总数，y 是给定 x_1,x_2,x_3,x_4,x_5 后的平均值，表示国家财政收入。其中的参数 β_0 称为回归常数，$\beta_1,\beta_2,\beta_3,\beta_4,\beta_5$ 称为回归系数，影响 y 的其他因素作用都包含在随机误差 ε 中，如果模型选择得合适，ε 应大致服从均值为 0 的正态分布。

直接利用 MATLAB 统计工具箱中的命令 regress 求解：得到模型的回归系数估计值及其置信区间（置信水平 α=0.05），检验统计量 R^2,F,p 的结果见表 3：

表 3 模型求解结果

参数	参数估计值	置信区间
β_0	82592	[29540,135640]
β_1	0.3148	[0.1518,0.4777]
β_2	-0.5955	[-1.2289,0.0380]
β_3	-0.3274	[-2.5138,1.8591]
β_4	0.5114	[0.2578,0.7651]
β_5	-0.7230	[-1.1974,-0.2486]
$R^2=0.9989$　　$F=3225.1$　　$p=0$		

2.3 结果的分析

（1）回归方程的显著性检验

$R^2=0.9989$ 指因变量 y 的 99.89%可由模型确定，$F=3225.1$ 远远超过 F 检验的临界值 finv(0.95,5,18)=2.7729，p 远小于 α，因而模型从整体来看是可用的。

（2）回归系数的显著性检验

β_2，β_3 的置信区间包含零点（但区间端点距零点很近），表明回归变量 x_2,x_3（对因变量 y 的影响）不是太显著的，但由时序残差图（图 3），残差均匀分布在 0 点线附近仍将 x_2,x_3 保留在模型中。

综上，建立的回归模型为：

$$y=82592+0.3148x_1-0.5955x_2-0.3274x_3+0.5114x_4-0.723x_5$$

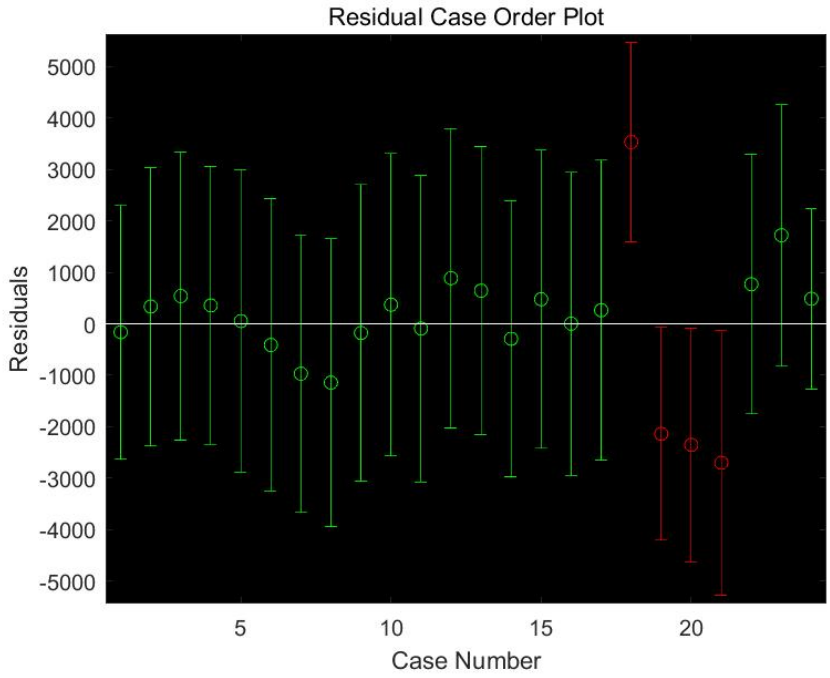


图 3 时序残差图

2.4 模型的改进

对模型进行改进，可以采用逐步回归模型：

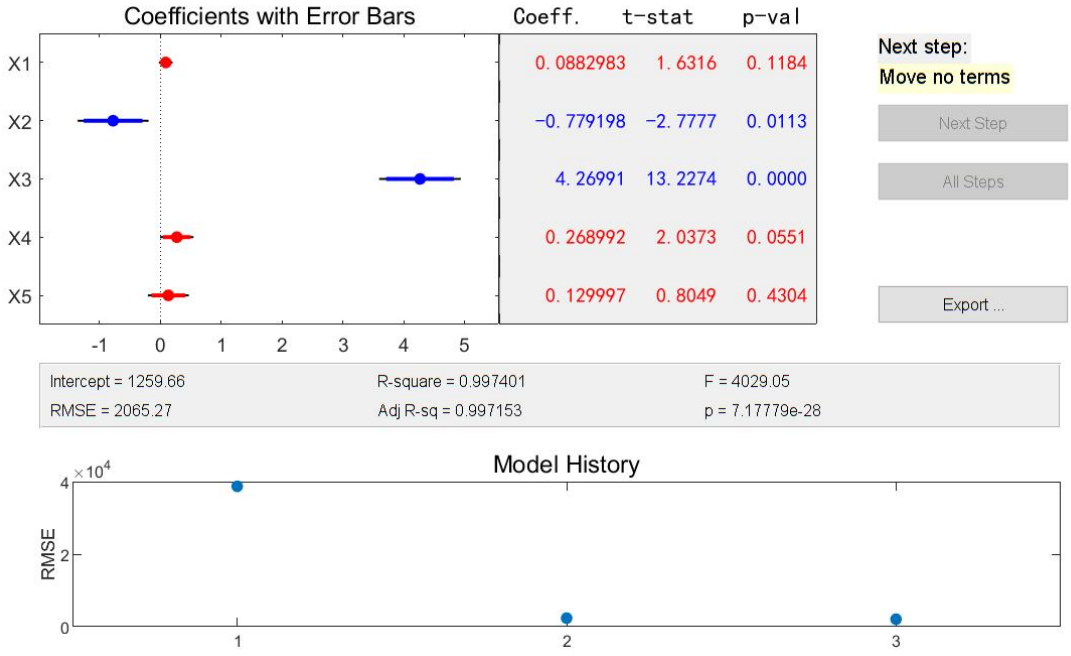


图 4 逐步回归结果

逐步回归模型直接把自变量 x_2, x_3 移除了模型，其他的自变量的相关系数分别为

0.088，0.268，0.129，截距为 1259.66。

3 多元线性回归版本 2

线性回归模型的向量化形式为：

$$\hat{y} = h_\theta(x) = \theta \cdot x$$

在此等式中：

- θ 是模型的参数向量，其中包含偏置项 θ_0 和特征权重 θ_1 至 θ_n。
- x 是实例的特征向量，包含从 $x^{(0)}$ 到 $x^{(n)}$，$x^{(0)}$ 始终等于 1。
- $\theta \cdot x$ 是向量 θ 和 x 的点积，它等于 $\theta_0 x^{(0)} + \theta_1 x^{(1)} + \theta_2 x^{(2)} + ... + \theta_n x^{(n)}$。
- h_θ 是假设函数，使用模型参数 θ。

使用均方误差作为模型的损失函数：

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\theta^T x_i - y_i)^2$$

- 最小二乘法求解析解

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\theta^T x_i - y_i)^2 = \frac{1}{m}[y_1 - \theta^T x_1, ..., y_m - \theta^T x_m]\begin{bmatrix} y_1 - \theta^T x_1 \\ ... \\ y_m - \theta^T x_m \end{bmatrix}$$

$$= \frac{1}{m}(y - X\theta)^T(y - X\theta) = \frac{1}{m}(y^T - \theta^T X^T)(y - X\theta)$$

$$= \frac{1}{m}(y^T y - y^T X\theta - \theta^T X^T y + \theta^T X^T X\theta)$$

其中 $y = \begin{bmatrix} y_1 \\ ... \\ y_m \end{bmatrix}, \theta = \begin{bmatrix} \theta_0 \\ ... \\ \theta_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1^T \\ ... & \\ 1 & x_m^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & ... & x_{1n} \\ ... & & & \\ 1 & x_{m1} & ... & x_{mn} \end{bmatrix}$。

接下来求导，导数为 0 的点即为使 MSE 最小的参数点：

$$\frac{\partial MSE}{\partial \theta} = \frac{1}{m}(0 - X^T y - X^T y + 2X^T X\theta) = 0 \leftrightarrow (X^T X\theta - X^T y) = 0 \leftrightarrow \theta = (X^T X)^{-1} X^T y$$

（矩阵求导参考：Petersen K B,Pedersen M S.The matrix cookbook[J].Technical University of Denmark,2008,7(15):510.）

- 批量（全）梯度下降 Batch Gradient Descent 求数值解

损失函数的梯度向量为：

$$\nabla_\theta MSE(\theta) = \begin{bmatrix} \frac{\partial MSE(\theta)}{\partial \theta_0} \\ ... \\ \frac{\partial MSE(\theta)}{\partial \theta_n} \end{bmatrix} = \frac{1}{m}(0 - X^T y - X^T y + 2X^T X\theta) = \frac{2}{m}X^T(X\theta - y)$$

梯度公式为：

$$\theta^{(next)} = \theta - \eta \nabla_\theta MSE(\theta)$$

其中，超参数η表示学习率，太低参数需要大量迭代损失函数才能收敛，太高参数可能越过山谷导致损失函数发散。

- 随机梯度下降 Stochastic Gradient Descent 求数值解

  梯度公式如上，但是训练集只有随机选择的一个实例。

- 小批量梯度下降 Mini-Batch Gradient Descent 求数值解

  梯度公式如上，但是训练集只有随机选择的一些实例。

表 算法比较

算法	核外支持	要求缩放	特点	Sklearn
标准方程	否	否	/	/
BGD	否	是	缺：每一步都是用整个数据集，如果数据集太庞大，算法变得很慢； 优：随特征数量扩展的表现比较好，如果拥有几十万个特征，比标准方程快得多	/
SGD	是	是	缺：永远定位不出最小值，解决方法是设置学习率调整函数，开始的步长比较大，然后越来越小（相当于抑制随机性） 优：可以逃离局部最优	SGDRegressor
MBGD	是	是	/	/

注：核外学习指的是可以处理计算机内存无法应对的大量数据，它将数据分成小批量，利用在线学习技术从小批量中学习。在线增量学习：指一个学习系统能不断地从新样本中学习新的知识，并能保存大部分以前已经学习到的知识。

要求缩放指的是对数据标准化处理，以加快训练速度。

模型参数	解释	默认值
fit_intercept	是否含有截距	True

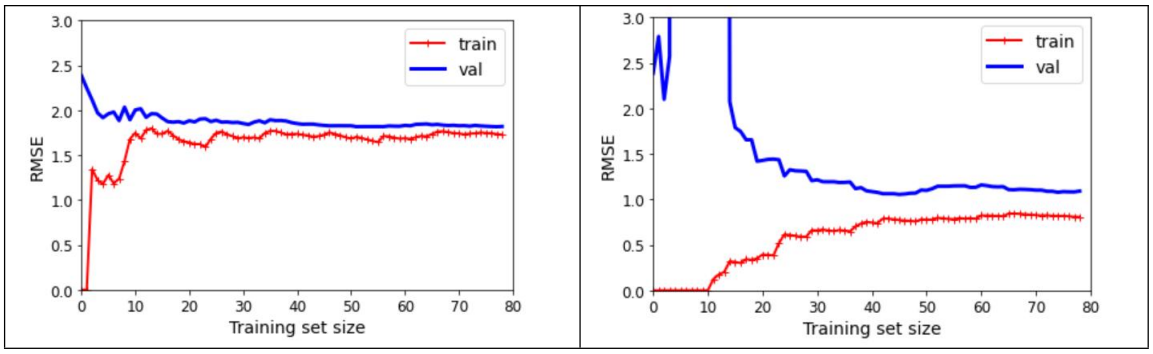
4 正则化

4.1 过拟合欠拟合

判断过拟合和欠拟合：

- 如果模型在训练集上表现良好，但根据交叉验证泛化较差，则模型过拟合；如果两者表现均不理想，则说明欠拟合。

- 观察学习曲线：模型在训练集和验证集上关于训练集大小（或迭代次数）的性能函数。



左图欠拟合，两条曲线接近且都很高：模型在训练集上的性能：当训练集只有一个或者两个实例时，模型可以很好地拟合他们，这是曲线从 0 开始的原因。随着新实例的加入，由于数据有噪声，另外选的模型过于简单，使得拟合效果差误差一直增大，直至平稳状态。在验证集上的性能：很少的实例训练后无法正确泛化，这是刚开始误差很大的原因。随着新实例的加入，模型学的更多，误差直至平稳状态。

右图过拟合，两条曲线间隙太大，正确的是验证误差达到训练误差。

缓解措施：

过拟合训练数据：简化模型，收集更多训练数据，减少噪声。（通过约束模型使其简单，降低过拟合的风险，这个过程叫做正则化，例如一元线性回归有两个参数 θ_0,θ_1，该算法在拟合数据时调整模型的自由度为 2，如果我们强行让 $\theta_1=0$，那么算法的自由度将会降为 1，并且拟合数据将变得更为艰难—能做的就是将水平线上移或者下移来尽量接近训练实例，最后极有可能停留在平均值附近。如果允许修改 θ_1 但是强制它只是很小的值，那么算法的自由度位于 1 和 2 之间，这个模型将会比自由度为 2 的模型稍微简单，但比自由度为 1 的模型稍微复杂，需要找到平衡点确保模型很好地泛化。在学习时，应用正则化的程度可以通过一个超参数来控制）

欠拟合训练数据：复杂模型，提供更好特征集。

4.2 岭回归等

在模型参数数量较大，而训练数据不够多的情况下，常用正则化缓解过拟合，对线性回归模型进行 l2 正则化得到岭回归：

$$J(\theta)=MSE(\theta)+\alpha\frac{1}{2}\sum_{i=1}^{n}\theta_i^2==MSE(\theta)+\alpha\frac{1}{2}\theta^2$$

$$\theta = (X^T X + \alpha A)^{-1} X^T y$$

其中，上式θ不含θ_0，下式A为$(n+1)*(n+1)$的单位阵，但左上角为0。

注：训练时将正则化项加入损失函数MSE，评估时用的不加正则化的MSE，损失函数与性能指标不同很常见，因为损失函数应该具有对优化友好的导数，而性能指标应该尽可能接近最终目标，例如交叉熵训练分类，而精确率召回率评估。

注：怎么选择超参数的值？做法之一是使用100个不同的超参数值来训练100个不同的模型，对测试集的测试误差进行比较，选择最小的（过拟合测试集）；做法之二是对验证集的误差进行比较，选择最小的，然后在完整的训练集（包括验证集）上训练得到最终模型。

模型参数	解释	默认值
alpha	乘以L2项的常数，控制正则化强度。	1
fit_intercept	是否含截距项	True
solver*{'auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 'saga', 'lbfgs'}*	求解方法	auto
random_state	随机种子	None

对线性回归进行l1正则化得到lasso回归：

$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$

注：Lasso回归的一个重要特点是它倾向于完全消除掉最不重要特征的权重。

模型参数	解释	默认值
alpha	乘以L2项的常数，控制正则化强度。	1
fit_intercept	是否含截距项	True

还有介于l2正则化与l1正则化之间的弹性网络：

$$J(\theta) = MSE(\theta) + r\alpha \sum_{i=1}^{n} |\theta_i| + \frac{1-r}{2} \alpha \sum_{i=1}^{n} \theta_i^2$$

5 非线性回归

非线性回归可以看成拟合非线性函数。

5.1 一元非线性回归

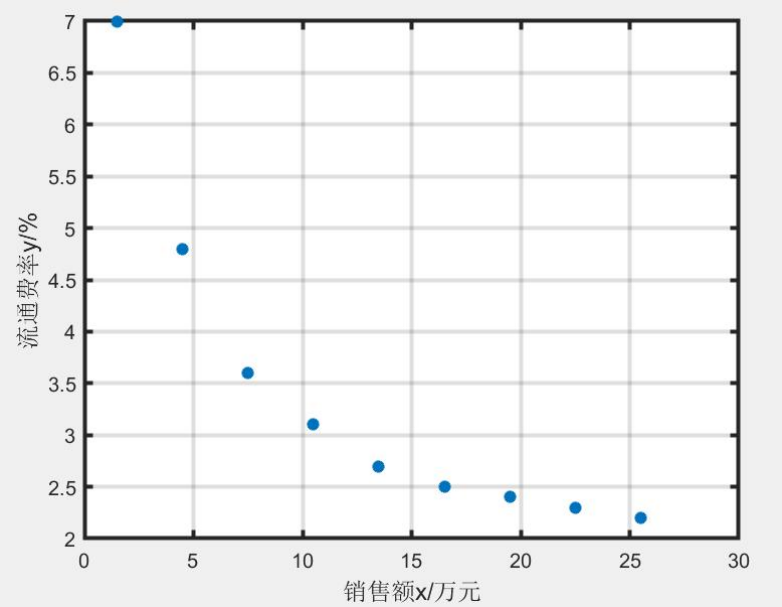
（1）一元非线性回归

销售	1.5	4.5	7.5	10.5	13.5	16.5	19.5	22.5	25.5

额									
流通费率	7.0	4.8	3.6	3.1	2.7	2.5	2.4	2.3	2.2

以销售额为自变量，流通费率为因变量研究非线性函数关系？

解答：首先绘制散点图观察函数形式：

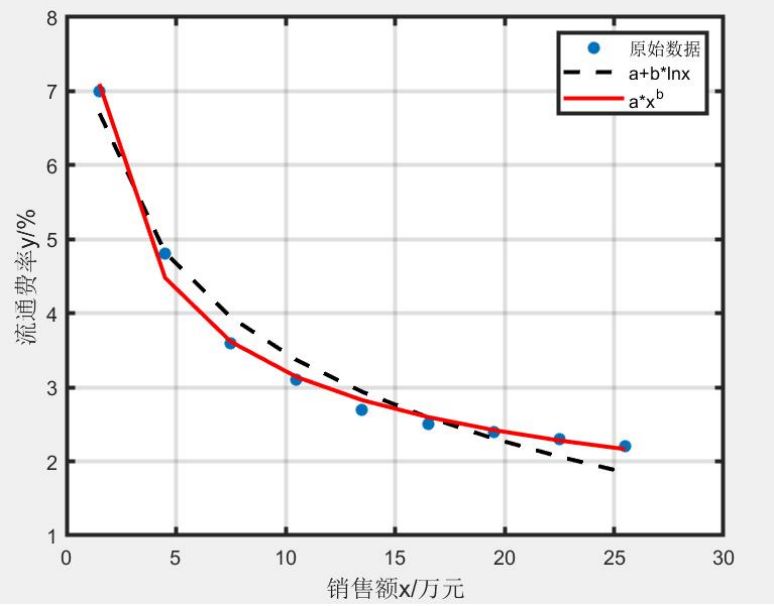


通过散点图可以看出可以采用对数函数 $a+b\cdot\ln x$ 或者幂函数 $a\cdot x^{b}$ 的形式，利用 Matlab 编程求得函数为：

$$y=7.3979-1.713\ln x$$

$$y=8.4112\cdot x^{-0.41893}$$

拟合图像为：



其中对数形式的拟合优度为 0.973，幂形式的拟合优度为 0.993，幂形式的函数

关系更符合。

5.2 多元非线性回归

y	x1	x2	x3
8.55	470	300	10
3.79	285	80	10
4.82	470	300	120
0.02	470	80	120
2.75	470	80	10
14.39	100	190	10
2.54	100	80	65
4.35	470	190	65
13.00	100	300	54
8.50	100	300	120
0.05	100	80	120
11.32	285	300	10
3.13	285	190	120

$$y=\frac{\beta_4 x_2-\dfrac{x_3}{\beta_5}}{1+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}$$

0.0627763848939352
0.0400481573042565
0.112415779863047
1.25259734178071
1.19136632668397

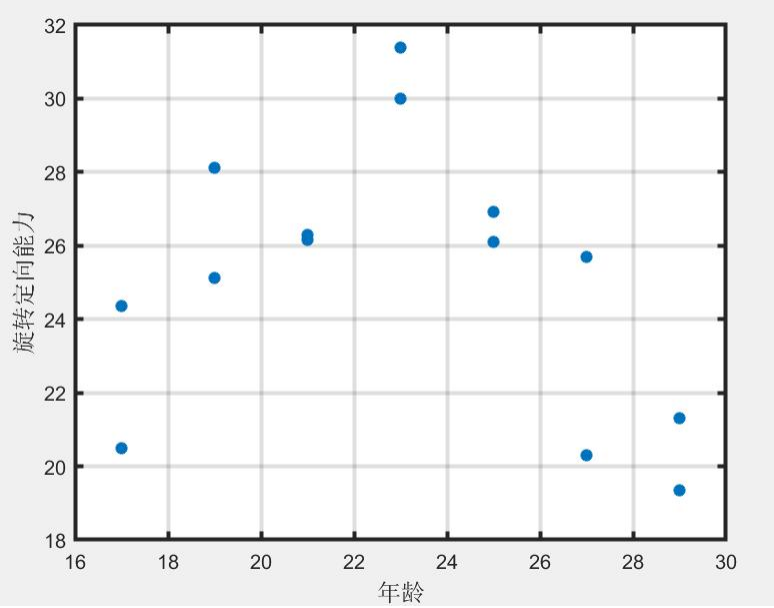
5.3 一元多项式回归

将 17 至 29 岁的运动员每两岁一组分为 7 组，每组两人测量其旋转定向能力，以考察年龄对这种运动能力的影响。现得到一组数据如下表所示：

年龄	17	19	21	23	25	27	29
第一人	20.48	25.13	26.15	30.0	26.1	20.3	19.35
第二人	24.35	28.11	26.3	31.4	26.92	25.7	21.3

以年龄为自变量，二人的旋转定向能力为因变量用一元多项式回归建立二者之间的关系？

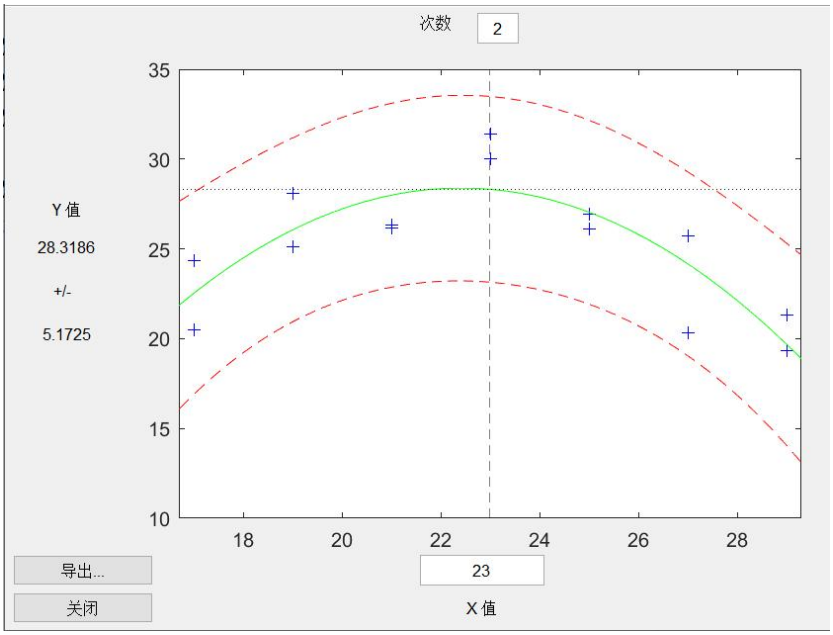
解答：首先绘制散点图观察函数形式，数据的散点图明显呈现两端低中间高的形状，所以应该拟合一条二次曲线。



选用二次模型 $y=a_2x^2+a_1x+a_0$，利用 matlab 编程求得模型为：

$$y=-0.2003x^2+8.9782x-72.2150$$

交互式画面如下图所示，给出 x 固定时 y 的曲线及其置信区间。



5.4 多元二项式回归

4.2.2　多元二项式回归

统计工具箱提供了一个作多元二项式回归的命令rstool，它也产生一个交互式画面，并输出有关信息，用法是

rstool(x,y,model,alpha)

其中输入数据x,y分别为$n\times m$矩阵和n维向量，alpha为显著性水平α（缺省时设定为0.05），model由下列4个模型中选择1个（用字符串输入，缺省时设定为线性模型）：

linear(线性)：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m$

purequadratic(纯二次)：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{j=1}^{m}\beta_{jj}x_j^2$

interaction（交叉）：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j\ne k\le m}\beta_{jk}x_jx_k$

quadratic(完全二次)：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j,k\le m}\beta_{jk}x_jx_k$

我们再作一遍例2 商品销售量与价格问题，选择纯二次模型，即

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{11}x_1^2+\beta_{22}x_2^2 \qquad (44)$$

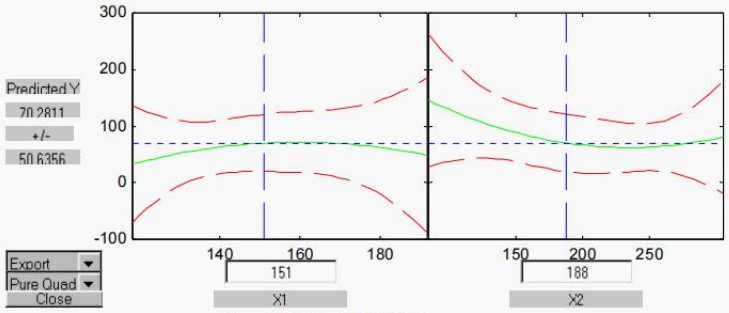


图2　拟合的交互式画面

得到一个如图2所示的交互式画面，左边是x_1（=151）固定时的曲线$y(x_1)$及其置信区间，右边是x_2（=188）固定时的曲线$y(x_2)$及其置信区间。用鼠标移动图中的十字线，或在图下方窗口内输入，可改变x_1,x_2。图左边给出y的预测值及其置信区间，就用这种画面可以回答例2提出的“若某市本厂产品售价160（元），竞争对手售价170（元），预测该市的销售量”问题。

图的左下方有两个下拉式菜单，一个菜单Export用以向Matlab工作区传送数据，包括beta(回归系数)，rmse（剩余标准差），residuals(残差)。模型（41）的回归系数和剩余标准差为

beta =−312. 5871　　7. 2701　　−1. 7337　　−0. 0228　　0. 0037

rmse =16.6436

另一个菜单model用以在上述4个模型中选择，你可以比较一下它们的剩余标准差，会发现以模型（24）的rmse=16.6436最小。

注意本例子在Matlab中完全二次模型的形式为

$$y=b_0+b_1x_1+b_2x_2+b_3x_1x_2+b_4x_1^2+b_5x_2^2 \qquad (45)$$