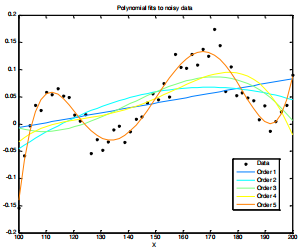
**拟合与多元线性回归**

1. **拟合多项式。**载入文件randomData.mat，这个文件中包括x和y两个变量。编写脚本对其做第一、第二、第三、第四和第五次多项式拟合。仿照下图，用蓝色的点绘制x和y的散点图，然后在同一坐标系下，用不同颜色绘制多项式拟合曲线，并在合适的位置加入标注(**xlabel,ylabel,legend**)。要想获得拟合预测的误差估计，你需要用到**polyfit**（[P,S,MU] = polyfit(X,Y,N)）的第3输出MU：x的均值MU(1)和标准差MU(2)，详见**help。**x的均值和标准差可以对数据标准化（XHAT = (X-MU(1))/MU(2)），使其均值为0，方差为1。这个标准化变换改进了多项式和拟合算法的数值特性。还需要使用与之搭配的**polyval**（[Y,DELTA] = polyval(P,X,S,MU)）,该函数需要输入**polyfit返回的**S和 MU来计算拟合误差DELTA。最终图像应如下图所示：



1. **经济增长模型**

增加生产、发展经济所依靠的主要因素有增加投资、增加劳动力以及技术革新等，在研究国民经济产值与这些因素的数量关系时，由于技术水平不像资金、劳动力那样容易定量化，作为初步的模型，可认为技术水平不变，只讨论产值和资金、劳动力之间的关系。在科学技术发展不快时，如资本主义经济发展的前期，这种模型是有意义的。

用*Q，K，L*分别表示产值、资金、劳动力，要寻求的数量关系*Q(K,L)*。经过简化假设与分析，在经济学中，推导出一个著名的Cobb-Douglas生产函数：

*Q(K,L) = aKαLβ*， 0<α,β<1 （\*）

式中α,β*，a*要由经济统计数据确定。现有美国马萨诸塞州1900—1926年上述三个经济指数的统计数据，如下表，试用数据拟合的方法，求出式（\*）中的参数α*,*β*，a*。

表1

|  |  |
| --- | --- |
| *t Q K L* | *t Q K L* |
| 1900 1.05 1.04 1.05  1901 1.18 1.06 1.08  1902 1.29 1.16 1.18  1903 1.30 1.22 1.22  1904 1.30 1.27 1.17  1905 1.42 1.37 1.30  1906 1.50 1.44 1.39  1907 1.52 1.53 1.47  1908 1.46 1.57 1.31  1909 1.60 2.05 1.43  1910 1.69 2.51 1.58  1911 1.81 2.63 1.59  1912 1.93 2.74 1.66  1913 1.95 2.82 1.68 | 1914 2.01 3.24 1.65  1915 2.00 3.24 1.62  1916 2.09 3.61 1.86  1917 1.96 4.10 1.93  1918 2.20 4.36 1.96  1919 2.12 4.77 1.95  1920 2.16 4.75 1.90  1921 2.08 4.54 1.58  1922 2.24 4.54 1.67  1923 2.56 4.58 1.82  1924 2.34 4.58 1.60  1925 2.45 4.58 1.61   1. 2.58 4.54 1.64 |

提示：由于（\*）式关于参数α,β*，a*是非线性的，因此，可以有两种方式进行拟合，一是直接使用MATLAB软件中的曲线或曲面拟合命令。另一个是将非线性函数转化成线性函数的形式，使用线性拟合。

1. **多元线性回归**。线性回归是一类应用广泛的统计模型，它用来拟合标量因变量y和一个或多个相互独立的自变量之间的关系。假设你对做了一个具有p个自变量的实验或观测，具有值，测得结果是一个实数y。重复n次，可以得到n组观测值, 组合得到以下矩阵：



线性回归的目标是通过找到合适的和b，拟合每一个与其对应的，使其符合以下n个方程：



换句话说，我们想建立如下的关系：



你需要对p=1的情况非常熟悉：找到最优的拟合曲线。特别，最小二乘法是最著名的方法。在本练习中，我们会用多种方式来实现最小二乘法，每种方法都有自己的优点，以及一些自然的推广。

1. 将文件夹中的regression.mat文件载入到工作空间
2. 文件中有数值矩阵X（数据）和Y（标签），在本例中，它们实际上是列向量（n=100，p=1）。
3. 当你有数据的时候，你应该做的第一件事就是数据可视化！幸运的是，给定的矩阵X和Y是一维的，所以二维散点图就足够了。将数据绘制为黑点。
4. 分析该图，数据大多显示出一种线性关系，并有一些区域异常值。现在试着用最小二乘法拟合一条直线。我们希望通过最小化误差的平方和（因此是“最小二乘法”），找到合适的和来拟合数据：



直觉上，如果每个都很接近，那说明我们已经找到了一个很好的拟合。我们将使用**fminsearch**通过数值优化来最小化目标函数。在这样做之前，为方便起见，我们把和合成一个向量。注意。通过将一列1合并到X，创建一个扩展数据矩阵XAug。

1. 回想一下**fminsearch**是如何使用的：我们要为它提供一个函数，该函数对任一，求目标函数的值。编写一个函数cost=squaredCost（beta，dataAug，labels）来计算的值，其中的参数为dataAug中给定的扩展数据矩阵和labels中的数据。

注意：不要使用XAug和Y！函数无法访问工作区中的变量；它们将通过squaredCost的输入参数导入。

选做：如果不用for循环可以实现该功能吗？优化速度会快得多。

1. 注意，我们的函数有三个输入参数，但是我们只希望最小化关于beta的目标函数，同时保持dataAug和labels作为参数，这些参数在我们的最小化过程中是恒定的。为了使用**fminsearch**，我们需要定义一个新的匿名函数，它只有一个参数（beta）。此函数将调用先前声明的squaredCost函数，并将工作区变量XAug和Y作为参数导入。完成以下命令以声明新的匿名函数squaredCostReduced=@（beta）squaredCost（beta，XAug，Y）
2. 现在通过使用SquaredCostReduced和初始值调用**fminsearch**。参阅**help**，看看前两个输入应该是什么。使用零作为初始值，注意它需要与大小相同.显示beta，并使用squaredCost（或squaredCostReduced）和beta计算成本。绘制具有数据点的拟合线（具有beta中的系数）进行验证。
3. 由于只有一个独立变量，我们也可以尝试用**polyfit**对数据进行拟合。调用**polyfit**来求X和Y的拟合直线，并验证返回的系数是否与**fminsearch**中的系数匹配（因为**polyfit**也是做最小二乘拟合）。
4. 可以证明，最小二乘拟合有一个解析解，称为正态方程（可以通过求导并令其等于零得到）：



验证这是否提供了相同的系数集（使用XAug）。

1. 事实上，因为我们有一个超定的线性方程组，所以我们可以用左除”\”来解。验证你的答案。
2. 用**regress**来估计参数，给出参数的区间估计，线性模型是否有效?
3. 这时，你可能认为我们浪费了时间多此一举。嗯，不完全是！首先，请注意**polyfit**不能处理多个独立变量。其次，解析解只适用于我们前面看到的二次目标函数，即误差平方之和。这个目标函数是在一个简单的概率模型假设条件下推出的，即采样输出Y是具有正态分布（高斯）噪声的线性关系。然而，这个噪声分布在许多情况下是不符合实际的，需要修改。看一下你绘制的曲线，你会发现拟合曲线由于一些离群值而不能很好地拟合大多数数据。实际上，用**绝对偏差误差代替平方和误差**，可以显著提高模型对异常值的鲁棒性（选做：为什么？这样的目标函数是什么意思？）:



重复第（d）-（f）部分，这次要编写一个不同的函数cost=absoluteCost（beta，dataAug，labels）。请注意，比较两个不同成本函数的值没有意义。将数据、最小二乘曲线（蓝色）和最小绝对偏差曲线（红色）绘制在一起。

这个目标函数的优化模型没有解析解！

1. 选做：探索目标函数使用其他p范数时会发生什么：



p=1的情况对应于最小绝对偏差，p=2对应于最小二乘。请修改你的Cost函数，把p也包括进去，这样你只需要写一个函数文件！