第四章作业

问题回顾 下表中的数据表示在液体环境下细菌的增长过程，寻找能体现这些数据趋势的最佳拟合方程，并预测 40 天后细菌的数量。（可以尝试线性拟合、二次拟合和指数拟合等）

天数	0	4	8	12	16	20
数量(10-6)	67	84	98	125	149	185

问题分析预测 40 天后细菌的数量可以建立多种不同的数学模型，这种模型可以依据于广泛研究的统计法则，也可以是依据于深入探讨的细菌繁殖机理模型，显然不同的方法预测结果也不太可能相同，并且各种方法也有各自的优越性以及局限性。本文以上两个方面展开，用定量化的手段（主要是通过置信区间以及假设检验的思想）来论述预测结果的好坏。

线性回归

对于存在关系的两组观测量，线性关系式最简单的一种，自然把模型建立在线性回归基础上也是最为自然的。

一般线性回归方程式为

$$\mathrm{Y} = \mathrm{a} + \mathrm{b} \cdot \mathrm{x} + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

根据最小二乘的原则，编写了如下基于 MATLAB 代码，脚本程序是为了使得代码调试更加方便。

而事实上线性回归的系数估计值的计算也不复杂，即：

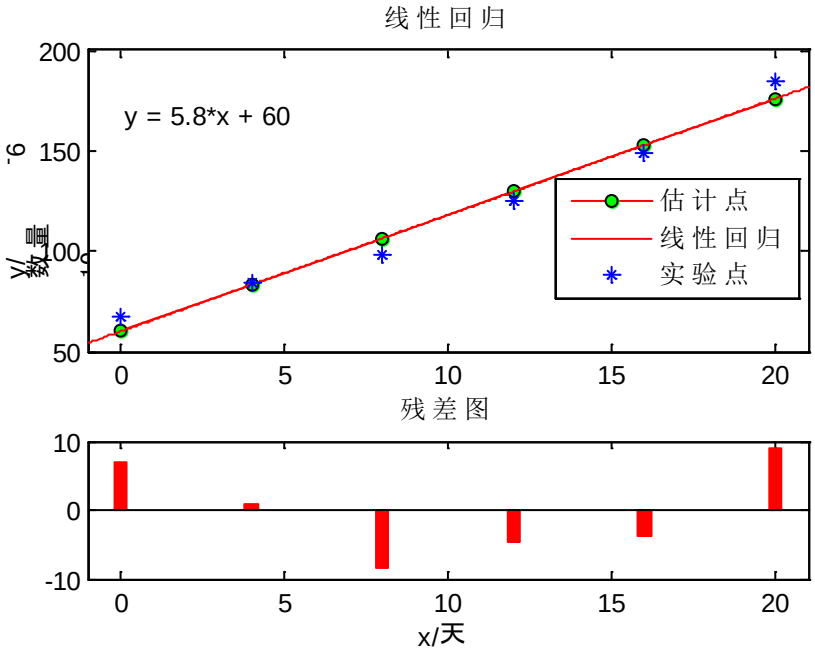
$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

是一本标准的二元线性方程组$Ax = b$

由于系数矩阵 A 形式固定，并且在这个特定的细菌数量试验中，一般研究人员按照实验要求等时间间隔地观测细菌数量，即 A 的数值一般来说都是固定的，只有右边系数矩阵随着实验对象的不同会有所变化，而本问题中 A 的条件数 cond(A)= 465.2574，可以直接用 A 的逆矩阵来求解。

```
%linear regression
n = length(x);
A1 = [n sum(x);sum(x) sum(x.^2)];
b1 = [sum(y); sum(x.*y)];
a = A1\b1;        % a is in increasing power series
a=rot90(rot90(a));  % let the polynomial arranged in decreasing powers.
y_es = polyval(a,x);
```

可以得到如下的数据分析（残差分析）图：



回归程度的数字特征

根据理论推导[1]关于线性回归有如下结论。

$$t = \frac{\hat{b}}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2)$$

由此可以做 b 的置信水平为$1-\alpha$ 的置信区间，区间构造如下：

$$|t| = \frac{|\hat{b}|}{\hat{\sigma}}\sqrt{S_{xx}} \le t_{\frac{\alpha}{2}}(n-2)$$

其中

$\hat{b}$为 b 的估计量

$$\widehat{\sigma^2} = \frac{Q_e}{n-2}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2$$

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

即回归的 b 值的置信区间为

$$\left(\hat{b} \pm t_{\frac{\alpha}{2}}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right)$$

根据以上结论编写程序如下

```
% check the significancy of the regression
Sxx=sum((x-mean(x)).^2);
Qe = sum((polyval(a,x)-y).^2);
alpha = 0.05;
delta = tinv(1-alpha/2 , n-2) *sqrt(Qe/(n-2)/Sxx);
```

[1] 盛骤，概率论与数理统计，高等教育出版社，第三版，303 页， 2007 年

```
[a(1)-delta, a(1)+delta]  % 95%回归参数（系数）置信区间
% estimation of x = 40
x0 = 40;
delta = tinv(1-alpha/2 , n-2) *sqrt(Qe/(n-2))...
    *sqrt(1+1/n+(x0-mean(x))^2/Sxx);
Y0 = polyval(a,x0);
[ Y0-delta, Y0+delta];  % 95%置信预测区间
```

本例子中的 95%的置信区间为

$$(4.5234, 7.0766)$$

该区间没有包含 0 点。从假设检验的角度来说，我们可以 95%的置信率，可以接受数据的线性关系假设。

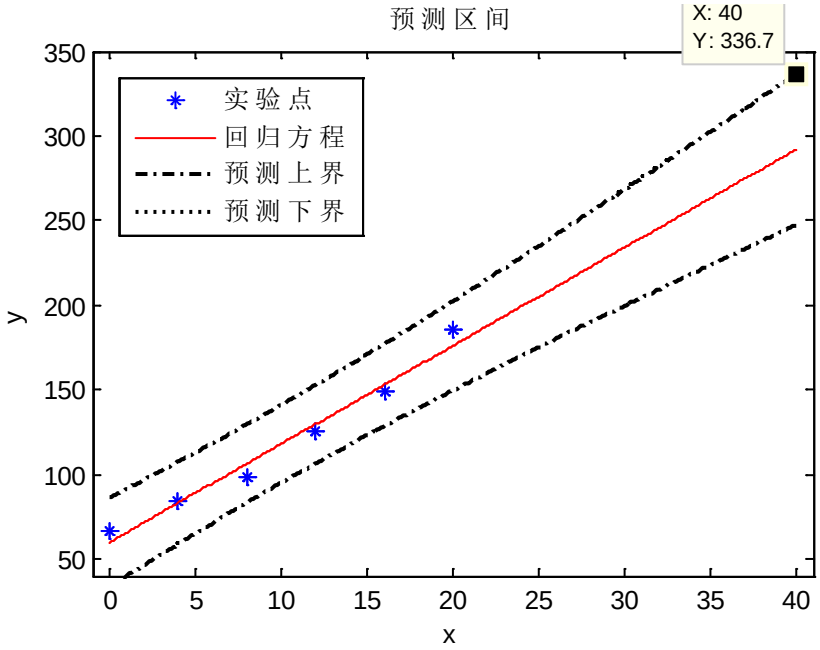
根据理论推导，Y 的观察值的点预测区间为：

$$\left(\hat{a}+\hat{b}x_0 \pm t_{\frac{\alpha}{2}}(n-2)\,\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right)$$

代入$x_0 = 40$ 得：

	Y0=a+bx0	95%预测区间	拟合优度R^2
x=40	292	(247.2870 , 336.7130)	0.9755

下面作图表示了 0~40 天的点预测，和预测区间，置信度为 95%。



从图上看线性回归的程度不错， b 的置信区间也比较小，但是预测区间却比较大，上下界相差了 30%。随着预测点原理实验数据区间，置信区间将变大。如果增大置信度，区间将变的更加大。

二次多项式回归

当一次回归的效果不显著时，可以采用二次多项式回归。当多项式回归方程次数较高的时候应该采用正交多项式做多项式拟合[2]，但如果次数不高可以考虑将二次多项式回归转化为一次多元多项式回归问题。

$$y = a_0 + a_1 x + a_2 x^2 + \epsilon$$

以上二次回归问题转化为

$$y = a_0 + a_1 x_1 + a_2 x_2 + \epsilon$$

其中$x_1 = x\,; x_2 = x^2$，而$\epsilon \sim N(0, \sigma^2)$。

回归方程，写成向量形式为

$$\hat{Y} = X \cdot \hat{\beta}$$

$$X_i = (1, x_{i1}, x_{i2})$$

$$\beta = (a_0, a_1, a_2)^{\mathrm{T}}$$

下面列出 n 个数据 k 阶次回归系数的计算方法，数学表示为：

$$\mathrm{X} = \begin{bmatrix} 1 & \mathrm{x}_1 & \cdots & \mathrm{x}_1^{\mathrm{k}} \\ 1 & \mathrm{x}_2 & \cdots & \mathrm{x}_2^{\mathrm{k}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{x}_{\mathrm{n}} & \cdots & \mathrm{x}_{\mathrm{n}}^{\mathrm{k}} \end{bmatrix};\ \mathrm{y} = \begin{bmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \vdots \\ \mathrm{y}_{\mathrm{n}} \end{bmatrix}$$

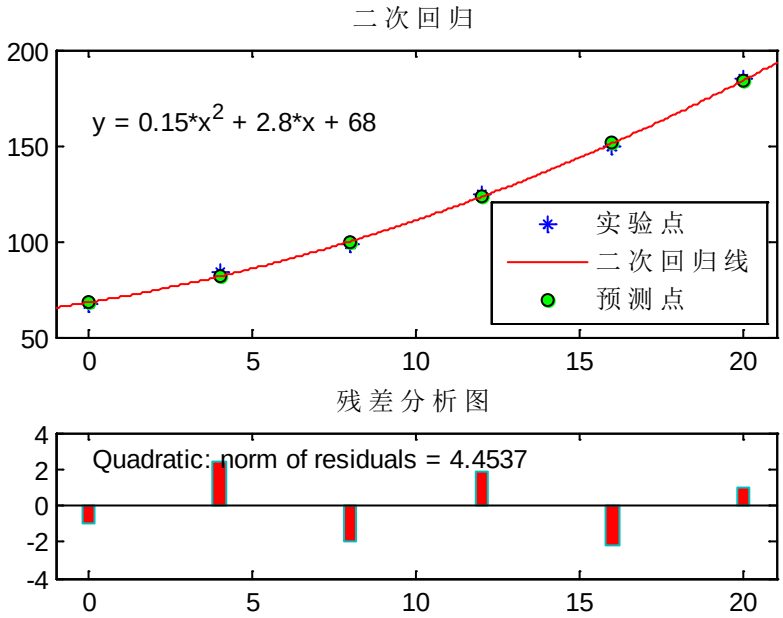
$$A = X^T X;\quad b = X^T Y$$

求解$A\beta = b$即可得到系数向量。下面给出程序

```
% k_order regression
X = ones(n,1);
for i=1:k
X = [X , x'.^i];  % k=1
end
A2 = X'*X;
b2 = X'*y';
a = A2\b2;
a=rot90(rot90(a));  % let the polynominal arranged in decreasing powers.
y_es = polyval(a,x);
```

回归结果如下图：

[2] http://202.121.199.249/foundrymate/lessons/data-analysis/24/243.htm



为了定量分析回归效果，与一元线性回归有类似，可以构造 $\beta_i = a_{i-1}$ 的统计量：

$$t = \frac{\widehat{\beta_i} - \beta_i}{\sqrt{c_{ii} \cdot \frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n-k-1}}} \sim t(n-k-1)$$

其中

$$C = (X^{'}X)^{-1}$$

$$c_{ii} = [C]_{ii}$$

k 为多元变量个数

则构造β_i的置信水平为α置信区间为：

$$\left(\widehat{\beta_i} \pm t_{\frac{\alpha}{2}}(n-k-1) \cdot \sqrt{c_{ii} \cdot \frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n-k-1}}\right)$$

根据上面的数学原理可以得到如下程序：

```
alpha = 0.05;  % 1-alpha是置信度
C = inv(A2);
delta = tinv(1-alpha/2 , n-k-1) *...   % 系数置信区间长度
   sqrt(diag(C).*sum((y_es-y).^2)/(n-k-1));
delta=rot90(rot90(delta));     % in increasing power series
```

得到各个参数的估计表

$y = a_0 + a_1x_1 + a_2x_2$	a_2	a_1	a_0
估计值	0.1507	2.7866	68.0357
95%置信下界	0.0670	1.0425	60.6190
95%置信上界	0.2344	4.5307	75.4524

同理可得任意观察点 x 的置信区间

$$\left(\hat{Y}_0 \pm t_{\frac{\alpha}{2}}(n-k-1)\cdot\sqrt{(1+X_0CX_0')\cdot\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{n-k-1}}\right)$$

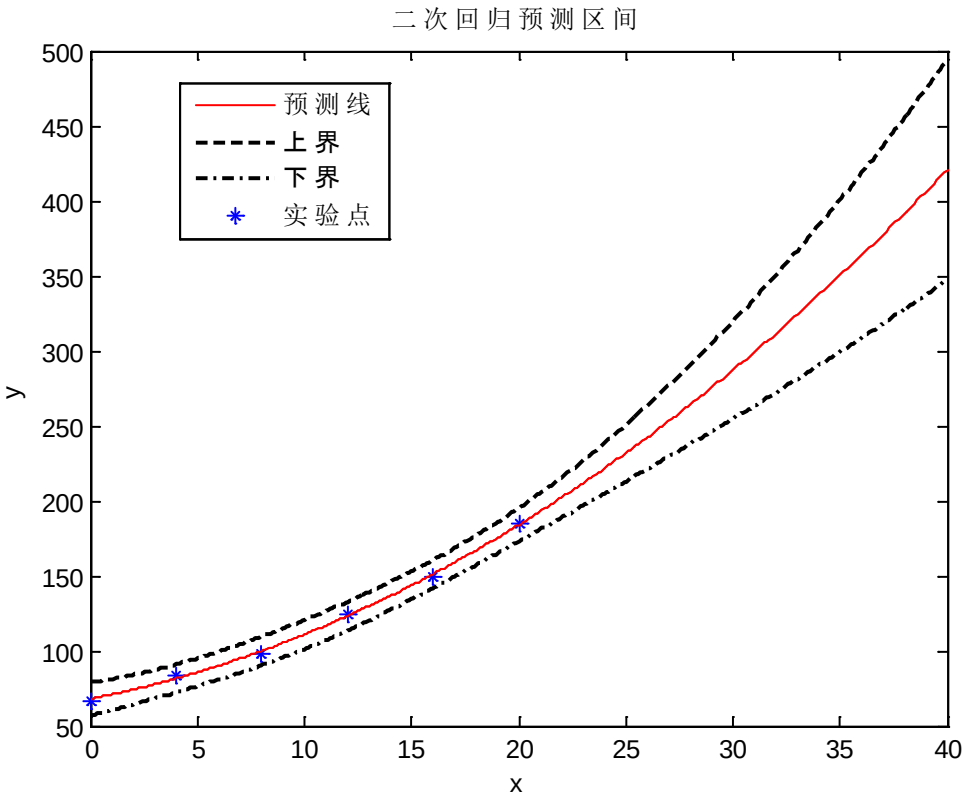
$$X_0=(1,x_0,x_0^2)'$$

预测区间计算程序为：

```
for i=1:k
X0 = [X0 , x0'.^i];  % X0为要预测的点
end
sigma = sum((y_es-y).^2)/(n-k-1);
delta = tinv(1-alpha/2 , n-k-1) *...
    sqrt((eye(401)+X0*C*X0').* sigma);
delta = diag(delta)';
```

得到 x0=40 时的区间为：

$$(347.1175\ ,494.0253)$$



从统计数据上看，二次回归的拟合优度$R^2=0.9979$ 较一次回归更好，但是预测的区间长度却变大了。

我们自然地想到要用更加高的阶次来拟合方程，比如用三阶多项式。秩序修改上面程序中的参数 k 就可以得到想要的结果（95%置信率）：

三阶多项式	a_3	a_2	a_1	a_0
估计值	0.0039	0.0335	3.6429	67.2857
95%置信下界	-0.0196	-0.6812	-2.1094	55.4400

95%置信上界	0.0274	0.7482	9.3952	79.1314

看到三次、二次、一次系数的置信区间都包含 0 点，所以这样的拟合也不理想，不能接受数据服从 3 次多项式关系的假设。

二次拟合的一般方法是将一个一元多项式的转化成了一个二元一次多项式，本文所采用的就是其中最为简单的一种，但是这种方法首先x_1与x_2 之间不能保证独立性，其次计算时向量$A = X'X$的条件数随着阶次的升高而升高，如果较高次的多项式就不适用这种方法。这时就要采用正交多项式的技巧来构造离散内积为 0 的多项式簇。

事实上由前面的 3 次多项式的拟合结果可以看到，更高次数的拟合意义不大，因为这样的拟合既没有明确的物理意义，也没有很好的拟合效果。可以看到从数学建模的角度，对细菌增长进行机理分析是十分有效的。

从数学模型出发——指数估计

细菌这样的生物是依靠细胞分裂来繁衍后代的，那么一种最朴素的模型就是，细菌的数量应当服从指数关系，下面就是一种可能的关系：

$$y = e^{a+b\cdot x+\epsilon}$$

$$\epsilon \sim N(0,\sigma^2)$$

这种关系可以十分容易地用前面的算法求解参数以及评价回归优度。

$$\ln y = a + b\cdot x + \epsilon$$

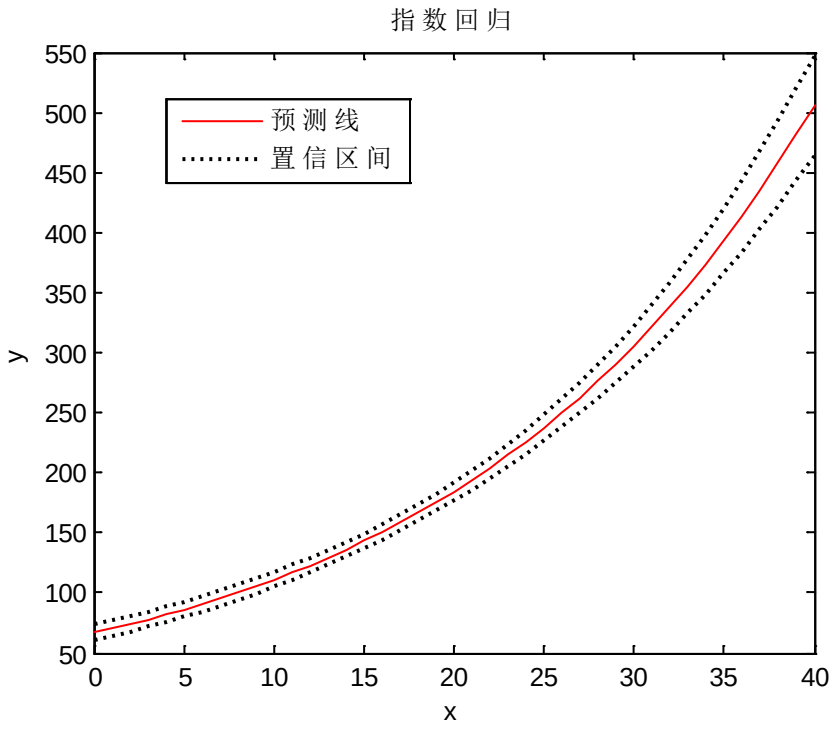
通过对数处理，方程又回到了前面的标准形式。

指数拟合 $y = e^{a+b\cdot x+\epsilon}$	b	a
估计值	0.0471	4.1701
95%置信下界	0.0535	4.2484
95%置信上界	0.0471	4.1701

拟合优度为 0.9983[3]。

指数回归的预测区间为：

[3] 不同的模型处理方法会有不同的处理结果，如果直接在指数函数的基础上使用最小二乘算法和取对数后使用最小二乘结果是不同的。



特别地x = 40的预测区间为：(463,548)。预测值为 y=505*10^-6。

可以看到同样置信水平下，指数回归模型兼有预测区间小，拟合优度好的优点，事实上这是由于指数回归模型从细菌增长的物理事实出发进行拟合，模型更为有效。

从数学模型出发——logistic 模型

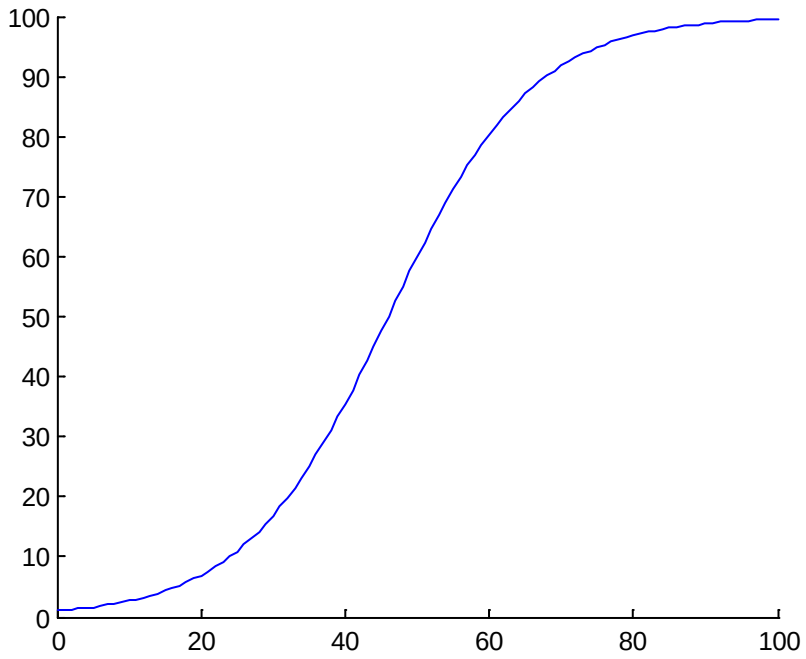
对于细菌这样的种群增长模型，已经有许多的研究，包括马尔科夫模型、马尔萨斯模型、宋健人口模型等等。这里要用的是一种成名已久的 Logistic 模型，也称作是阻滞增长模型。模型的方程式：

$$\lambda \frac{dy}{dx} = (y_{max} - y) \cdot y$$

这是一个非线性微分方程，它的解析解为：

$$y = \frac{y_{max}}{1 + \left(\frac{y_{max}}{y(0)} - 1\right) e^{-\lambda x}}$$

下图所示就是一个典型的 logistic 增长模式：



由于该方程高度非线性，难以线性化处理。故用 MATLAB 中的 cftool 工具箱进行拟合。结果如下:

General model:

f(x) = a/(1+(a/67-1)*exp(b*x))

Coefficients (with 95% confidence bounds):

a = 1e+004 (fixed at bound)

b = -0.05118 (-0.0522, -0.05016)

Goodness of fit:

SSE: 17.03

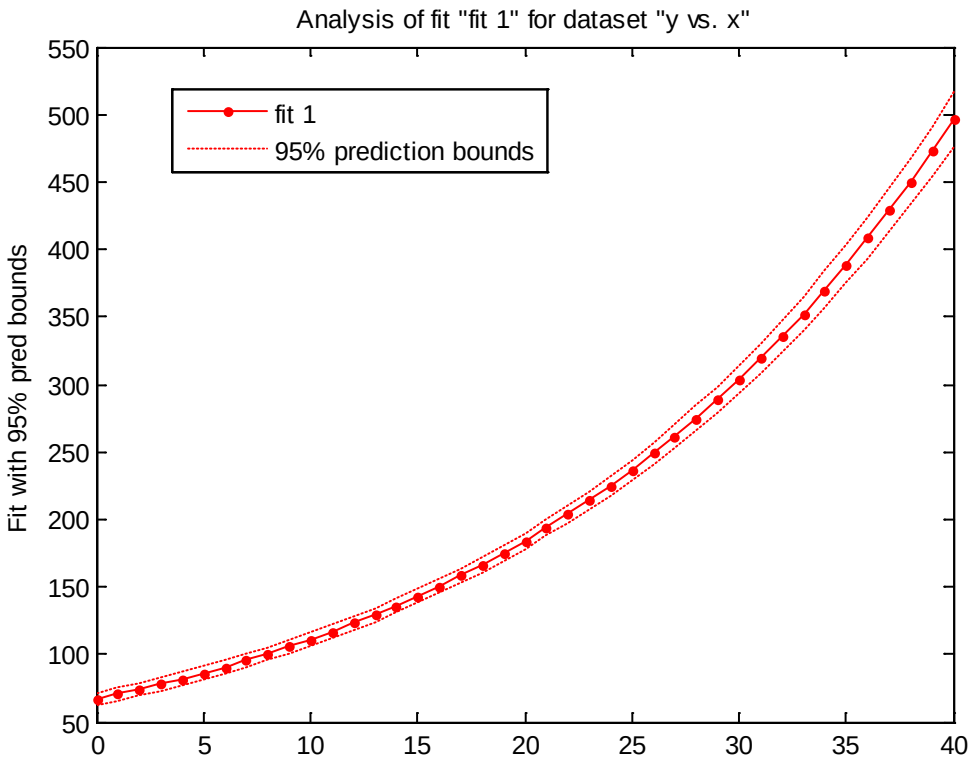
R-square: 0.9982

Adjusted R-square: 0.9982

RMSE: 1.845

可以看到高亮部分，即y_{max}总是达到模型计算的预制边界，但有趣的是拟合效果确实非常好。这是由于我们实验的数据主要集中在 logistic 模型的上升阶段，这个时候细菌生长的环境资源充沛阻力小，近似为前面的指数增长。也就是说这里使用 Logistic 模型有些大材小用了。

如果假设细菌生长的极限值为$y_{max} = 10^4$：那么用 cftool 得到的预测区间如下图：



特别地，$x = 40$时，区间为(476.689,516.375) 预测值是 496.532 基本上与前面的指数区间重合。

通过 Logistic 模型的计算，我们强烈建议对实验中的培养基测定一个极限细菌值：

$$y_{max} = y(t \to \infty)$$

这样就可以使用 logsitic 模型了。

综上我们可以认为实验 40 天后细菌的数量将大概达到 450~500 左右，有 95%的置信度。

插值方法

由于本题的背景取自细菌数量测定，细菌的数量 y 的实验测定本省就带有随机误差，也就是前面模型中反复使用的ϵ ，并且注意到研究人员每隔固定的天数观测细菌数量，观测的时间间隔是否严格相等并不能保证。从模型机理的角度说，本问题并不适宜用差值的方法来进行函数的预测，因为数据真值并不一定经过实验点，特别是对于预测点离实验区间比较远的情况。

小结

本问探讨了细菌预测问题的几个模型，并进行了求解以及分析了模型的预测能力。通过实际操作发现，对于这样一个有着深刻生物繁殖机理背景的问题，一般需要采用较好的数学模型，这里就表现为应该使用 logistic 模型。根据模型对细菌的繁殖提出了更有意义的实验设计要求，既要测量长时间后的细菌停止增长的数量，这样整个模型就更有效了。