第三章
回归预测法

金　林
中南财经政法大学统计系
jinlin82@qq.com

2016 年春

Outline

1 线性回归预测法
- 模型
- 参数估计
- 模型检验
- 预测

2 非线性回归预测法

3 应用回归预测法应注意的问题

- 模型
- 参数估计
- 模型检验
- 预测

模型及其假设

1. 一元线性回归模型：
$$y_i = b_0 + b_1 x_i + u_i$$
其中，b_0, b_1 是未知参数，u_i 为剩余残差项或称随机扰动项。
2. u_i 满足一定的假设条件：
   1. u_i 是一个随机变量；
   2. u_i 的均值为零；
   3. 在每一个时期中，u_i 的方差为常量，即 $D(u_i) = \sigma_u^2$ ；
   4. 各个 u_i 相互独立；
   5. u_i 与自变量无关。

- 模型
- **参数估计**
- 模型检验
- 预测

最小二乘法

1. 线性回归模型常见参数估计方法有普通最小二乘法和最大似然估价法。
2. 最小二乘法的基本思想是使得残差平方和最小。
3. 用最小二乘法进行参数估计，得到的估计表达式为：

$$\hat{b}_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}$$

- 模型
- 参数估计
- 模型检验
- 预测

标准误差

1. 标准误差：估计值与因变量值间的平均平方误差。
2. 其计算公式为：

$$SE = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

可决系数

1. 可决系数：衡量自变量与因变量关系密切程度的指标，表示自变量解释了因变量变动的百分比。
2. 其计算公式为：
$$R^2 = SSR/SST$$
其中 $SSR = \sum(\hat{y} - \bar{y})^2$ ，$SST = \sum(y - \bar{y})^2$
3. 可决系数取值于 0 与 1 之间，并取决于回归模型所解释的 y 方差的百分比。

相关系数

1. 其计算公式为：

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

2. 可决系数是相关系数的平方。
3. 相关系数越接近 $+1$ 或 -1，两个变量之间的相关关系越密切。
4. 如果直线从左至右上升，则相关系数为正；
5. 如果直线从左至右下降，则相关系数为负。

相关系数与可决系数之间的区别

1. 相关系数测定变量之间的密切程度，可决系数测定自变量对因变量的解释程度。
2. 相关系数有正负，可决系数只有正号。正相关系数意味着因变量与自变量以相同的方向增减。

回归系数显著性检验

1. 自变量回归系数的检验
2. 检验统计量为

$$t_b = \frac{\hat{b}}{S_{\hat{b}}}$$

其中：$S_{\hat{b}} = SE/\sqrt{\sum(x-\bar{x})^2}$，$t_b$ 服从自由度为 $n-2$ 的 t 分布。

3. 取显著性水平为 α ，如果 $|t_b| > t_\alpha$ ，则回归系数 b 显著。

F 检验

1. 整体线性关系显著性检验
2. 三个平方和：SST、SSR 和 SSE
3. 其自由度分别为 $n-1$ ，自变量的个数 p 和 $n-p-1$ 。
4. 检验统计量

$$F = \frac{MSR}{MSE}$$

其中 $MSR = SSR/p$ ，$MSE = SSE/(n-p-1)$ 。

自相关的检验：DW 统计量

1. 什么是自相关
2. DW 检验:

$$DW = \frac{\sum_{i=2}^{n}(\hat{u}_i - \hat{u}_{i-1})^2}{\sum_{i=1}^{n}\hat{u}_i^2}$$

3. DW 统计量与一阶自相关系数 ρ 的关系：

$$DW \approx 2(1-\rho)$$

4. 决策准则

异方差检验:White 检验

1. 什么是异方差
2. White 检验：
    1. 估计方差 σ^2
    2. 以估计的 $\hat{\sigma}^2$ 作为因变量，以自变量二项式项作为自变量作辅助回归
    3. 得到辅助回归的判决系数 R^2 ，计算检验统计量 nR^2
    4. 当模型中不存在异方差时，$nR^2 \sim \chi^2(p)$ ，其中 p 为辅助回归中包含截距项在内的自变量的个数。
    5. 给定显著性水平 α, 若 $nR^2 > \chi^2_\alpha$, 则表明模型中存在异方差。

- 模型
- 参数估计
- 模型检验
- **预测**

点估计

1. 将新的自变量的值代入拟合的模型中

区间估计

1. 区分为因变量的平均值预测和因变量具体值预测
2. 有小样本和大样本的情况

因变量的平均值预测

1. 小样本：预测区间 $\hat{y} \pm t_{\alpha/2}(n-2)SE_{\hat{\mu}}$
2. 大样本：预测区间 $\hat{y} \pm z_{\alpha/2}SE_{\hat{\mu}}$
3. 其中

$$SE_{\hat{\mu}} = SE\sqrt{\frac{1}{n} + \frac{(x_F - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

在多元情况下，矩阵表达式为

$$SE_{\hat{\mu}} = SE\sqrt{\boldsymbol{X_F}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'_{\boldsymbol{F}}}$$

因变量的个别值预测

1. 小样本：预测区间 $\hat{y} \pm t_{\alpha/2}(n-2)SE_{\hat{y}}$
2. 大样本：预测区间 $\hat{y} \pm z_{\alpha/2}SE_{\hat{y}}$
3. 其中

$$SE_{\hat{y}} = SE\sqrt{1 + \frac{1}{n} + \frac{(x_F - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

在多元情况下，矩阵表达式为

$$SE_{\hat{y}} = SE\sqrt{\mathbf{1} + \boldsymbol{X_F}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_{\boldsymbol{F}}'}$$

1. 线性回归预测法
2. 非线性回归预测法
   - 常见曲线
   - 曲线回归参数估计
   - 曲线回归模型评价
   - 曲线的选择
3. 应用回归预测法应注意的问题

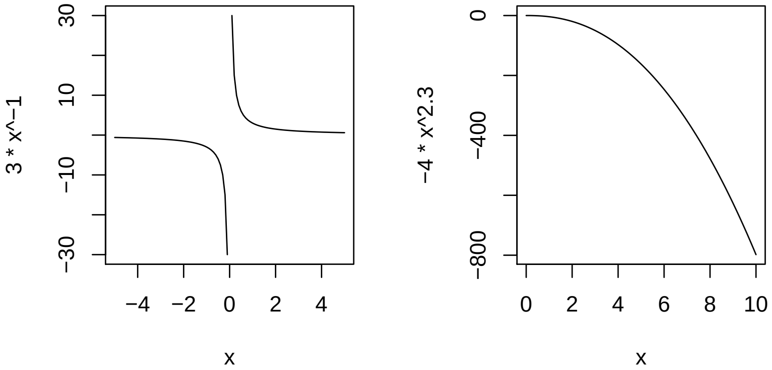
概述

1. 线性回归
2. 可以转化为直线回归的曲线回归 (变量变换)
3. 不能转化为直线回归的曲线回归
4. 非参数、半参 5 数回归
5. 神经网络、支持向量机

- 常见曲线
- 曲线回归参数估计
- 曲线回归模型评价
- 曲线的选择

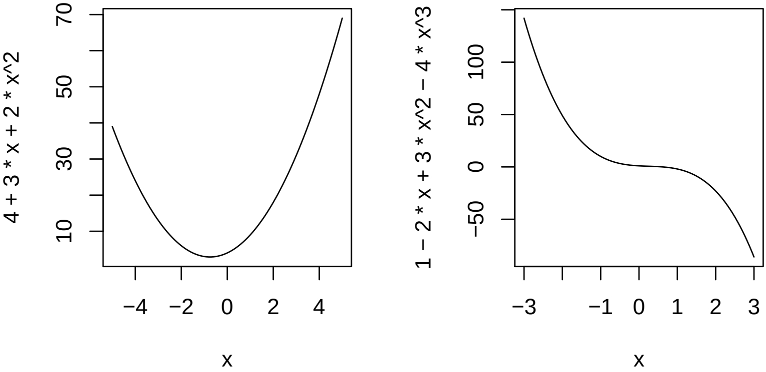
幂函数曲线

1. $f(x) = ax^b$



多项式曲线

1. $f(x) = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n$



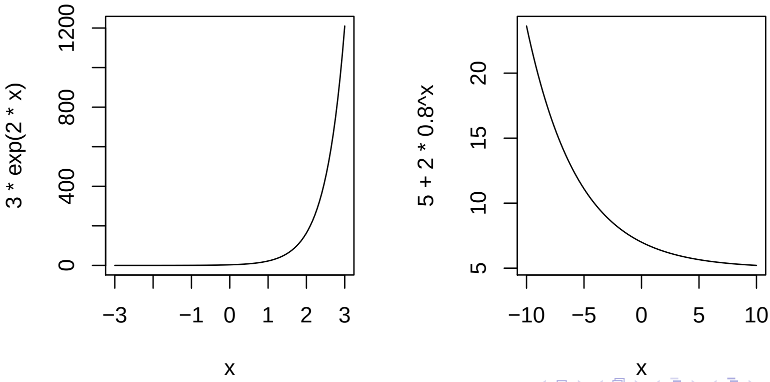
指数曲线

1. 指数曲线

$$f(x) = ae^{bx}$$

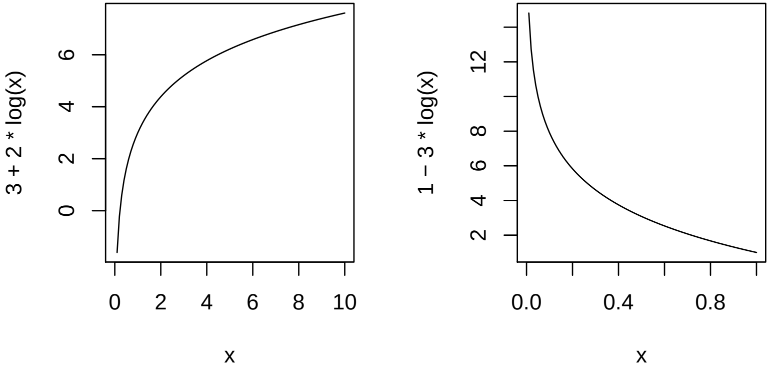
2. 修正指数曲线

$$f(x) = a + bc^{x}$$



对数曲线

1. $f(x) = a + b\ln x$



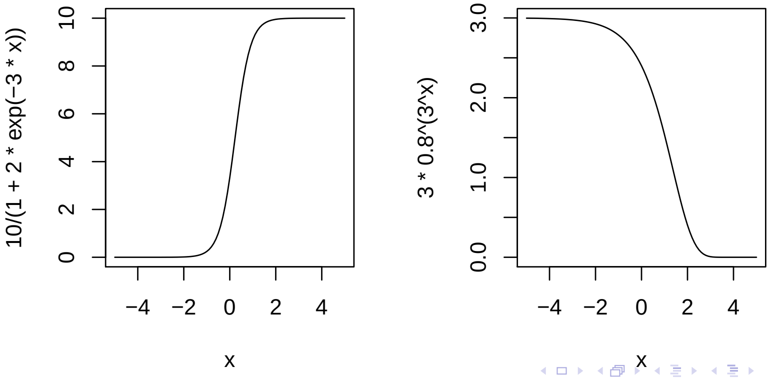
S 型曲线 (生长曲线)

1. 皮尔曲线

$$f(x) = \frac{L}{1 + ae^{-bx}}$$

2. 龚泊兹曲线

$$f(x) = ka^{b^x}$$



- 常见曲线
- 曲线回归参数估计
- 曲线回归模型评价
- 曲线的选择

最小二乘法

1. the nonlinear regression model:

$$y_i = f(x_i, \beta) + \varepsilon_i$$

   the function f is nonlinear.

2. minimisation of the residual sums of squares (RSS) with respect to β,

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2$$

3. the solution to the minimisation problem is the least-squares parameter estimates, which we denote $\hat{\beta}$.

数值方法

1. In contrast to linear regression, the minimisation of RSS will in general be a nonlinear problem due to the nonlinearity of f.
2. numerical optimisation methods are needed.
3. These methods are iterative procedures that will ideally approach the optimal parameter values in a stepwise manner.
4. At each step, the algorithms determine the new parameter values based on the data, the model, and the current parameter values.
5. the most common algorithm for estimation in nonlinear regression is the Gauss-Newton method, which relies on linear approximations to the nonlinear mean function at each step.

高斯 -牛顿算法

1. Given m functions $\boldsymbol{r} = (r_1, \cdots, r_m)$ of n variables $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)$, with $m \geq n$, the Gauss-Newton algorithm iteratively finds the minimum of the sum of squares：

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} r_i(\boldsymbol{\beta})^2$$

2. Starting with an initial guess $\boldsymbol{\beta}^{(0)}$ for the minimum, the method proceeds by the iterations

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \left(\mathbf{J_r}^T \mathbf{J_r}\right)^{-1} \mathbf{J_r}^T \mathbf{r}(\boldsymbol{\beta}^{(s)})$$

3. where, if $\boldsymbol{r}$ and $\boldsymbol{\beta}$ are column vectors, the entries of the Jacobian matrix are

$$(\mathbf{J_r})_{ij} = \frac{\partial r_i(\boldsymbol{\beta}^{(s)})}{\partial \beta_j}$$

and the symbol ^{T} denotes the matrix transpose.

数据拟合

1. In data fitting, where the goal is to find the parameters $\boldsymbol{\beta}$ such that a given model function $y = f(x, \boldsymbol{\beta})$ best fits some data points (x_i, y_i), the functions r_i are the residuals

$$r_i(\boldsymbol{\beta}) = y_i - f(x_i, \boldsymbol{\beta})$$

2. Then, the Gauss-Newton method can be expressed in terms of the Jacobian $\mathbf{J_f}$ of the function f as

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} + \left(\mathbf{J_f}^{\mathbf{T}}\mathbf{J_f}\right)^{-1}\mathbf{J_f}^{\mathbf{T}}\mathbf{r}(\boldsymbol{\beta}^{(s)})$$

3. The assumption $m \geq n$ in the algorithm statement is necessary, as otherwise the matrix $\mathbf{Jr}^{\mathbf{T}}\mathbf{Jr}$ is not invertible and the normal equations cannot be solved (at least uniquely).

例子

1.In this example, the Gauss-Newton algorithm will be used to fit a model to some data by minimizing the sum of squares of errors between the data and model's predictions. the data are in the following table.

i	x	y
1	0.038	0.050
2	0.194	0.127
3	0.425	0.094
4	0.626	0.2122
5	1.253	0.2729
6	2.500	0.2665
7	3.740	0.3317

例子

1. It is desired to find a curve (model function) of the form

$$y = \frac{\beta_1 x}{\beta_2 + x}$$

2. that fits best the data in the least squares sense, with the parameters β_1 and β_2 to be determined.
3. We will find β_1 and β_2 such that the sum of squares of the residuals

$$r_i = y_i - \frac{\beta_1 x_i}{\beta_2 + x_i}, (i = 1, \ldots, 7)$$

is minimized.

4. The Jacobian $\mathbf{J_r}$ of the vector of residuals r_i in respect to the unknowns β_j is an 7×2 matrix with the i-th row having the entries

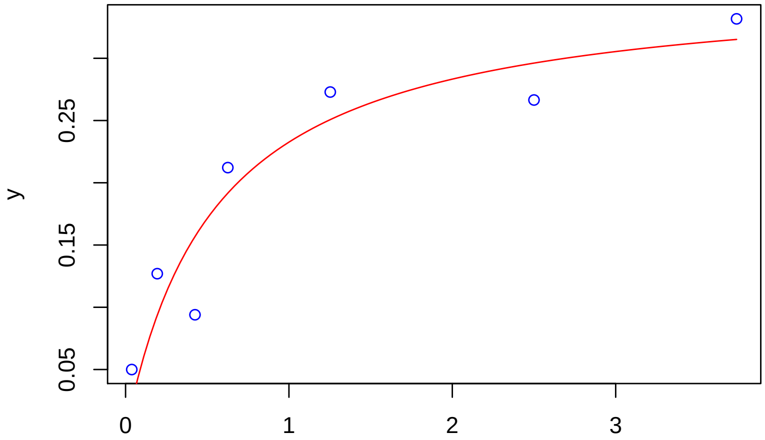
$$\frac{\partial r_i}{\partial \beta_1} = -\frac{x_i}{\beta_2 + x_i}, \ \frac{\partial r_i}{\partial \beta_2} = \frac{\beta_1 x_i}{(\beta_2 + x_i)^2}.$$

例子

1. Starting with the initial estimates of $\beta_1 = 0.9$ and $\beta_2 = 0.2$
2. after five iterations of the Gauss-Newton algorithm the optimal values $\hat{\beta}_1 = 0.362$ and $\hat{\beta}_2 = 0.556$ are obtained.
3. The sum of squares of residuals decreased from the initial value of 1.445 to 0.00784 after the fifth iteration.

例子图形

1. The plot in the figure shows the curve determined by the model for the optimal parameters versus the observed data.



数值方法存在的困难

1. how to start the procedure and how to choose the initial/starting parameter value.
2. how to ensure that the procedure reached the global minimum rather than a local minimum.
3. These two issues are interrelated.
4. it is very important to provide sensible starting parameter values.
5. Poorly chosen starting values will often lead the procedures astray so no useful model fit is obtained.
6. If lack of convergence persists regardless of the choice of starting values, then it typically indicates that the model in its present form is not appropriate for the data at hand.

非线性回归参数估计的特点

1. As the solutions to nonlinear regression problems are numeric, they may differ as a consequence of different algorithms, different implementations of the same algorithm , different parameterisations, or different starting values.
2. the resulting parameter estimates often will not differ much.
3. If there are large discrepancies, then it may possibly indicate that a simpler model should be preferred.

- 常见曲线
- 曲线回归参数估计
- 曲线回归模型评价
- 曲线的选择

模型评价与诊断

1. 类似线性回归模型
2. 判决系数 R^2
3. 均方误差
4. 残差图，残差分析

- 常见曲线
- 曲线回归参数估计
- 曲线回归模型评价
- 曲线的选择

模型曲线选择方法

1. 事前选择方法
    1. 根据过去理论或者经验
    2. 根据数据散点图的分布形状
2. 事后方法
    1. 曲线模型比较：包括参数意义和模型的拟合优度

注意的问题

1. 应用回归预测法时，应首先确定变量之间是否存在相关关系。
2. 如果变量之间不存在相关关系，对这些变量应用回归预测法就会得出错误的结果。
3. 正确应用回归分析预测时应注意：
    1. 用定性分析判断现象之间的依存关系；
    2. 避免回归预测的任意外推；
    3. 应用合适的数据资料。