MATH363 Week 3

Examples of Simple Linear Regression

> SLR是形如下式的线性模型：
>
> $$Y_i = a + bz_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$$
>
> 其矩阵表示如下：
>
> $$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \cdots & \cdots \\ 1 & z_n \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$
>
> 这就是 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 的矩阵表示形式
>
> 注意到，
>
> $$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_i z_i \\ \sum_i z_i & \sum_i z_i^2 \end{pmatrix}$$

我们不妨有如下表示：

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

$$S_{zz} = \sum_{i=1}^{n} z_i^2 - n\bar{z}^2 = \sum_{i=1}^{n}(z_i - \bar{z})^2$$

事实上，我们一般用这种表示方法来表示协方差，这里没有考虑自由度的原因是因为我们仅仅用这种表示方法来简化表达计算式

因此，

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{nS_{zz}}\begin{pmatrix} \sum_i z_i^2 & -\sum_i z_i \\ -\sum_i z_i & n \end{pmatrix} = \frac{1}{S_{zz}}\begin{pmatrix} \frac{\sum_i z_i^2}{n} & -\bar{z} \\ -\bar{z} & 1 \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_i y_i \\ \sum_i y_i z_i \end{pmatrix}$$

所以，

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \frac{1}{S_{zz}}\begin{pmatrix} \bar{y}\sum_i z_i^2 - \bar{z}\sum_i y_i z_i \\ \sum_i y_i z_i - n\bar{z}\bar{y} \end{pmatrix}$$

我们惊奇地发现，

$$\begin{cases} \hat{a} = \bar{y} - \hat{b}\bar{z} \\ \hat{b} = \frac{S_{yz}}{S_{zz}} \end{cases}$$

它们都是 a 和 b 的无偏估计，也就是说，$E(\hat{a}) = a$, $E(\hat{b}) = b$

估计值 $\hat{\boldsymbol{\beta}}$ 的方差为：

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma^2}{S_{zz}}\begin{pmatrix} \frac{\sum_i z_i^2}{n} & -\bar{z} \\ -\bar{z} & 1 \end{pmatrix}$$

根据这个方差-协方差矩阵我们不难发现，

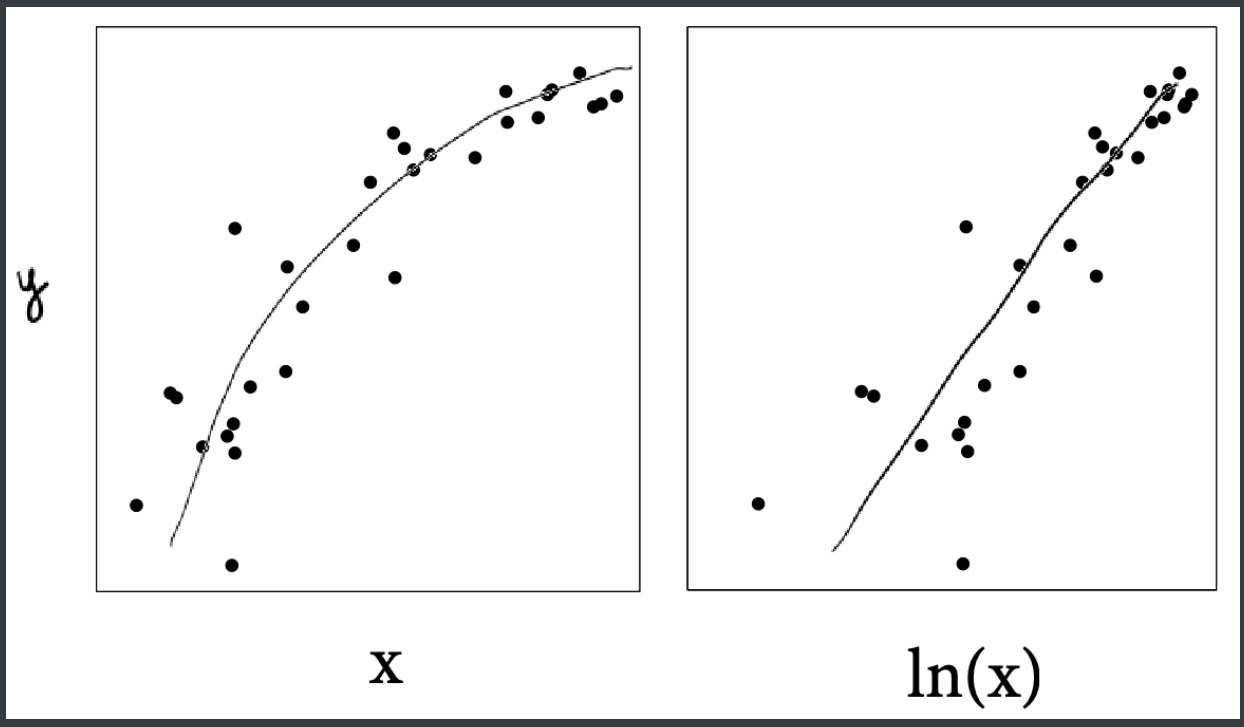
$$\begin{cases} Var(\hat{a}) = \frac{\sigma^2}{nS_{zz}} \sum_i z_i^2 = \sigma^2(\frac{1}{n} + \frac{\bar{z}^2}{S_{zz}}) \\ Var(\hat{b}) = \frac{\sigma^2}{S_{zz}} \end{cases}$$

并且可以得到估计值之间的协方差：

$$Cov(\hat{a}, \hat{b}) = -\frac{\bar{z}\sigma^2}{S_{zz}}$$

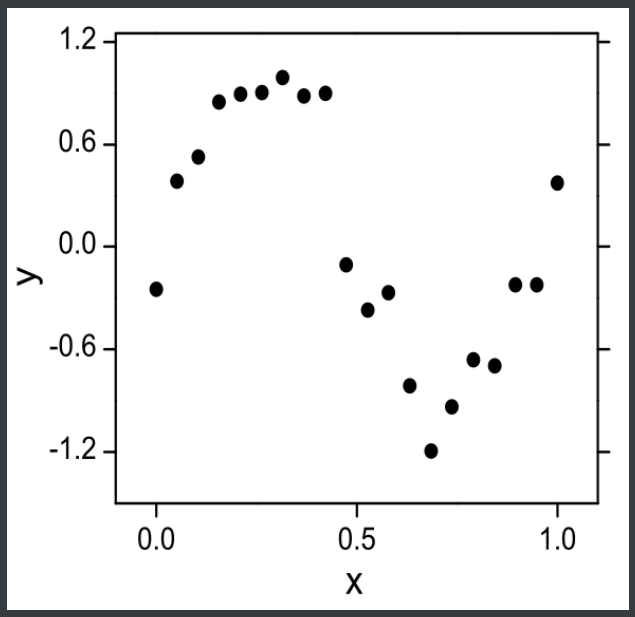
Plots

画图可以直观地显示模型是否合适，有时候适当变换一下variables可以让模型看起来更具有线性性质，如课件给出的例子 x 与 $ln(x)$：



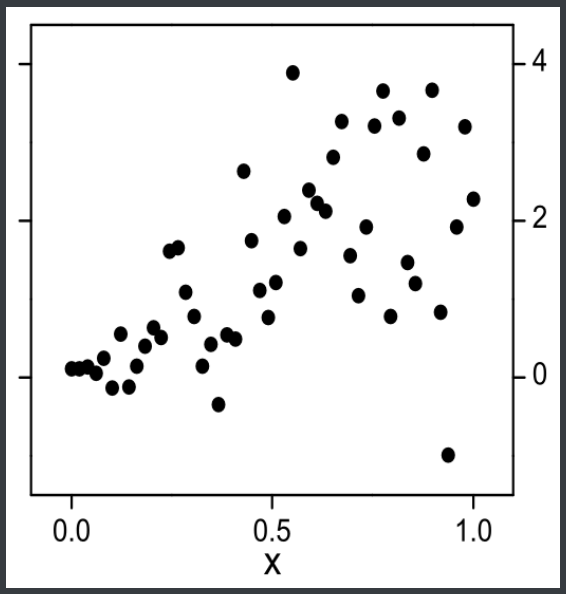
画图有时候也可以体现模型的一些问题

例如课件给出的左图：



1. SLR显然不适用因为无法用一条直线来“适当”地拟合
2. 如果强行假设SLR使用，首先这个拟合的直线应该接近于一条水平直线，这代表着因变量与自变量之间没有关系

又例如课件给出的右图：



若SLR适用，我们发现随着 x 的增长，y 愈加发散，也就是说 y 的方差越来越大，这显然与我们SLR的假设--$Var(y) = \sigma^2$ 相矛盾

Rank of Matrix

> 在用最小二乘法推导GLM中系数 β 的估计量的时候，我们假设了 $(p \times p)$ 矩阵 $\mathbf{X}^T\mathbf{X}$ 的逆矩阵存在，也就是说这个矩阵是可逆的，非奇异的；逆矩阵存在当且仅当这个矩阵是满秩的

> Definition of the Rank of Matrix
>
> Rank(X) = 行秩（线性无关行数）= 列秩（线性无关列数）$\leq \min\{n, p\}$

Orthogonal Structure 正交结构

> 如果两个向量 x_1 和 x_2 是正交的，那么 $x_1^T x_2 = 0$

Design Matrix 中的正交结构

> 将 $(n \times p)$ 的矩阵 X 分为 $(x \times q)$ 的矩阵 X_1 和 $(n \times (p-q))$ 的矩阵 X_2, 也就是说 $X = [X_1, X_2]$, 使得 $X_1^T X_2 = 0$ （即分解为两个正交部分），那么对于最小二乘估计量：
>
> $$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$
>
> In detail,
>
> $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 可以被改写为：
>
> $$\mathbf{Y} = X_1\beta_1 + X_2\beta_2 + \varepsilon$$
>
> 注意到：
>
> $$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix} \begin{pmatrix} X_1 & X_2 \end{pmatrix} = \begin{pmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{pmatrix}$$
>
> $$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} (X_1^T X_1)^{-1} & 0 \\ 0 & (X_2^T X_2)^{-1} \end{pmatrix}$$

> 于是，
>
> $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y} = \begin{pmatrix} (X_1^T X_1)^{-1} X_1^T \mathbf{y} \\ (X_2^T X_2)^{-1} X_2^T \mathbf{y} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$
>
> 我们惊奇地发现，$\hat{\beta}_1$ 只依赖于 X_1，$\hat{\beta}_2$ 只依赖于 X_2
>
> 如果我们将 β_2 从模型中移除（相当于 $\beta_2 = 0$），模型就对应地变成 $\mathbf{Y} = X_1\beta_1 + \varepsilon$，$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T \mathbf{y}$，可以发现这并不影响对 β_1 的估计
>
> 正交性Orthogonality意味着所施加的一些未知参数为0的条件并不影响对其他参数的估计
>
> 特别地，如果 X 的所有列都互相正交，那么 $\mathbf{X}^T\mathbf{X}$ 就是一个对角矩阵并且对于 β_j 的估计仅仅只依赖于 x_{ij}，也就是 j-th covariate那一列的观测值

Polynomial Regression 多项式回归

> 形如：
>
> $$Y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \ldots + \beta_p z_i^p + \varepsilon_i$$
>
> 那么，上式的Design Matrix就为：
>
> $$\begin{pmatrix} 1 & z_1 & z_1^2 & \ldots & z_1^p \\ 1 & z_2 & z_2^2 & \ldots & z_2^p \\ .. & .. & .. & .. & .. \\ 1 & z_n & z_n^2 & \ldots & z_n^p \end{pmatrix}$$
>
> 显然，各列不可能正交，这意味着一些参数的改变非常可能会改变其它参数估计值的量，这也同时意味着我们不能单独解释系数

Orthogonal Polynomials 正交多项式

> 为了避免上述这种dependence的情况，我们一般用如下的正交多项式（以三次cubic多项式为例）：
>
> $$Y_i = \beta_0 + \beta_1 \xi_{i1} + \beta_2 \xi_{i2} + \beta_3 \xi_{i3} + \varepsilon_i$$
>
> 其中，
>
> $\xi_{i1} = \xi_1(z_i)$ 是关于 z_i 的一次线性函数，也就是说 $\xi_1(z) = a_0 + a_1 z$
>
> $\xi_{i2} = \xi_2(z_i)$ 是关于 z_i 的二次函数，也就是说 $\xi_2(z) = b_0 + b_1 z + b_2 z^2$
>
> $\xi_{i3} = \xi_3(z_i)$ 是关于 z_i 的三次函数，也就是说 $\xi_3(z_i) = c_0 + c_1 z + c_2 z^2 + c_3 z^3$
>
> 上述函数中的系数 a, b, c 对于每一个特定的数据集都是固定fixed已知的，但不唯一not unique
>
> 总之，多项式列是关于观测值 x 的函数，各列之间需要正交，也就是说：
>
> $$Col_i^T Col_j = 0$$

考点补充

判断矩阵（Design Matrix X）是否满秩 full rank：

> 1. 行列式判断：$det(X) \neq 0$
>
> 行列式的计算：行列式等于矩阵任意一行或一列的元素 a_{ij} 与：
>
> $$A_{ij} = (-1)^{(i+j)} M_{ij}$$
>
> 的积的和

> 举个例子：求如下矩阵的行列式
>
> $$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$
>
> $$det(A) = a(-1)^{(1+1)}M_{11} + b(-1)^{(1+2)}M_{12} + (-1)^{(1+3)}M_{13}$$
>
> 也就是说，
>
> $$det(A) = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$
>
> 2. 用肉眼看矩阵各列之间是否存在线性相关

判断矩阵（Design Matrix X）是否为正交

> 看 X^TX 是否为对角矩阵，因为正交列的积为0