第二章：一元线性回归

主讲人：黄磊

数学学院
西南交通大学

September 3, 2019

Outline

1. 一元线性回归模型
   - 实际背景与理论模型
   - 参数估计
   - 最小二乘估计的性质
   - 回归方程的显著性检验
   - 残差分析
   - 回归系数的区间估计
   - 预测和控制
   - 本章总结

一元回归

需要研究某一现象和它的主要因素的关系，但又存在其他不确定因素，因此这是一种不确定关系。那么，如何刻画这种关系，如何估计、检验这种关系。

直观一些，我们可以画$\{x_i, y_i, i = 1, \ldots, n\}$的散点图, 在R里面可以用"plot()"，查看R自带函数，用"help(plot)"

Description

Generic function for plotting of R objects. For more details about the graphica

For simple scatter plots, plot.default will be used. However, there are p
methods(plot) and the documentation for these.

Usage

```
plot(x, y, ...)
```

Arguments

x the coordinates of points in the plot. Alternatively, a single plotting str

y the y coordinates of points in the plot, *optional* if x is an appropriate s

... Arguments to be passed to methods, such as graphical parameters (

type

一元线性回归模型的数学形式

只有一个解释变量x, 模型形式很简单

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{1.1}$$

这称为一元线性理论回归模型。有样本数据$\{(x_i, y_i), i = 1, \ldots, n\}$时，可得样本回归模型

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1.2}$$

(1.1)和(1.2) 实际是等价的，统称为一元线性回归模型。通常，还需要一下假设条件

1. 解释变量x的非随机性
2. 随机变量ε_i满足Gauss-Markov 条件，即零期望，等方差
3. $\{\varepsilon_i, i = 1, \ldots, n\}$ 的相互独立性

一元线性回归模型的数学形式

由此，响应变量y_i的期望和方差

$$E(y_i) = \beta_0 + \beta_1 x_i, var(y_i) = \sigma^2, i = 1, \ldots, n$$

接下来的首要任务就是，估计β_0和β_1，一旦有了估计值$\hat{\beta}_0, \hat{\beta}_1$, 就可得到一元线性经验回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

$\hat{\beta}_0, \hat{\beta}_1$的实际意义?。进一步为了区间估计和假设检验，还需假定(1.1)中的ε服从正态分布，

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

同理，

$$\varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$$

因此给定x_i时，y_i服从均值为$\beta_0 + \beta_1 x_i$方差为σ^2的正态分布，
即$y_i \sim N(\beta_0 + \beta_1 x, \sigma^2), i = 1, \ldots, n$

补充：矩阵形式

矩阵是处理线性关系的有力工具，今后会遇到各种矩阵或向量形式的表达，例如

$$
\begin{aligned}
\boldsymbol{y} &= (y_1, \ldots, y_n)^\top, \quad \boldsymbol{1} = (1, \ldots, 1)^\top \\
\boldsymbol{x} &= (x_1, \ldots, x_n)^\top, \quad \boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{x})_{n\times 2} \\
\boldsymbol{\varepsilon} &= (\varepsilon_1, \ldots, \varepsilon_n)^\top, \quad \boldsymbol{\beta} = (\beta_0, \beta_1)^\top
\end{aligned}
$$

其中，$\boldsymbol{X}^\top$即矩阵的转置，有时候也用$\boldsymbol{X}'$。于是模型(1.1)就变成了

$$
\begin{cases}
\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon \\
E(\varepsilon) = \boldsymbol{0} \\
Var\varepsilon = \sigma^2 \boldsymbol{I}_n
\end{cases}
\tag{1.3}
$$

其中$\boldsymbol{I}_n$就是n阶单位阵，R里面可以用$diag(n)$函数生成。

β的估计—LSE

Ordinary least square estimation, 普通最小二乘估计(LSE是一所好大学)。LSE考虑让观测值y_i与其回归值$E(y_i)=\beta_0+\beta_1 x_i$的离差越小越好，综合考虑样本数据的n个离差值，定义离差平方和如下，

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 \tag{1.4}$$

最小二乘估计，就是寻找$\hat{\boldsymbol{\beta}}=(\hat{\beta}_0,\hat{\beta}_1)^\top$, 以其使得$Q(\boldsymbol{\beta})$取得最小值，即

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in R^2} Q(\boldsymbol{\beta})$$

称$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$为$y_i$的回归拟合值，$e_i=y_i-\hat{y}_i$为残差。(1.4)中代入$\hat{\boldsymbol{\beta}}$之后，就叫做残差平方和。

$\hat{\beta}$的推导

关于(1.4)的最小化问题，由于$Q(\beta)$是关于β的非负二次函数，最小值总是存在。对偏导方程组求解，

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} & = \_\_\_\_\_\_ = 0 \\ \frac{\partial Q}{\partial \beta_1} & = \_\_\_\_\_\_ = 0 \end{cases} \tag{1.5}$$

整理得

$$\begin{cases} \beta_0 + \bar{x}\beta_1 & = \bar{y} \\ \bar{x}\beta_0 + \frac{\sum_{i=1}^n x_i^2}{n}\beta_1 & = \frac{\sum_{i=1}^n x_i y_i}{n} \end{cases} \tag{1.6}$$

其中，$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$, 或者$\bar{x} = \frac{1}{n}\mathbf{1}'\boldsymbol{x}$。解方程组(1.6), 得到最小二乘估计，标记为$(\hat{\beta}_0, \hat{\beta}_1)$。令$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$, 则

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{1.7}$$

$\hat{\beta}$的推导

思考$(\hat{\beta}_0, \hat{\beta}_1)$的最小性、唯一性？

考察Hessian矩阵，

$$\frac{\partial^2 Q}{\partial\beta\partial\beta^\top} = \begin{pmatrix} 2n & n\bar{x} \\ n\bar{x} & 2\sum_{i=1}^n x_i^2 \end{pmatrix} = 2\boldsymbol{X}'\boldsymbol{X},$$

只要$rank(\boldsymbol{X}) = 2$, 那么$\boldsymbol{X}'\boldsymbol{X}$ 正定的，$(\hat{\beta}_0, \hat{\beta}_1)$就是最小值，且唯一。

由(1.5)可以得残差的性质，

1. 残差平均为零，$\frac{1}{n}\mathbf{1}'\boldsymbol{e} = \frac{1}{n}\sum_{i=1}^n e_i = 0$
2. n维空间中的$\boldsymbol{e}$向量与$\boldsymbol{x}$正交，$<\boldsymbol{e}, \boldsymbol{x}> = \sum_{i=1}^n x_i e_i = 0$

分位数回归(不必掌握，扩展知识)

当(1.4)中的离差平方，变成了离差绝对值时，就是中位数(0.5分位数)回归了

$$Q(\beta) = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

当然更一般地，还有其他分位数$0 < \tau < 1$回归，

$$Q(\beta) = \sum_{\eta_i \geq 0} \tau |y_i - \beta_0 - \beta_1 x_i| + \sum_{\eta_i < 0} (1 - \tau) |y_i - \beta_0 - \beta_1 x_i|$$

where $\eta_i = y_i - \beta_0 - \beta_1 x_i$. 当$\tau = 0.5$的时候，怎样？

MLE

当$\varepsilon_i \sim N(0, \sigma^2)$, 可知$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, 得$y_i$密度函数

$$\frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\},$$

于是样本数据$\{y_i, i = 1, \ldots, n\}$的似然函数为

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\{-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\}$$

对数化$' - \ln'$, 记为$\ell(\boldsymbol{\beta}, \sigma^2)$, 变成最小化$\ell(\boldsymbol{\beta}, \sigma^2)$问题，

课堂练习：请同学们证明最小化$\ell(\boldsymbol{\beta}, \sigma^2)$与最小化$Q(\boldsymbol{\beta})$的等价性。

线性与无偏性

一、线性的定义：$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^\top$ 是关于随机变量$\{y_i, i = 1, \ldots, n\}$的线性函数，称之为线性估计量。由(1.7)得

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n c_i y_i = \boldsymbol{c}' \boldsymbol{y} \tag{1.8}$$

其中$c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$。课堂练习：同学们完成$\hat{\beta}_0$是线性估计量的推导。

二、无偏性的定义：对任意一个关于β的估计量$\hat{\beta}$，如果$E(\hat{\beta}) = \beta$, 则称$\hat{\beta}$为无偏估计量。

$$E(\hat{\beta}_1) = \underline{\hspace{8cm}} = \beta_1, \tag{1.9}$$

课堂练习：同学们完成$\hat{\beta}_0$的证明。 进一步就有，$E(\hat{y}) = \beta_0 + \beta_1 x_i$。

最小二乘估计量的方差

随机向量的预备知识

Lemma 1.1

若$Var(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n$, 且$\boldsymbol{c} = (c_1, c_2, \ldots, c_n)' \in R^n$, 则$\boldsymbol{c}'\boldsymbol{y}$的方差具有如下形式，

$$Var(\boldsymbol{c}'\boldsymbol{y}) = \sigma^2 \boldsymbol{c}'\boldsymbol{I}_n\boldsymbol{c} = \sigma^2 \boldsymbol{c}'\boldsymbol{c}.$$

若$\boldsymbol{c}_1, \boldsymbol{c}_2 \in R^n$, 则有协方差公式如下，

$$Cov(\boldsymbol{c}_1'\boldsymbol{y}, \boldsymbol{c}_2'\boldsymbol{y}) = \sigma^2 \boldsymbol{c}_1'\boldsymbol{I}_n\boldsymbol{c}_2 = \sigma^2 \boldsymbol{c}_1'\boldsymbol{c}_2$$

由公式(1.8), 以及Lemma 1.1得，

$$Var(\hat{\beta}_1) = \sigma^2 \boldsymbol{c}'\boldsymbol{I}_n\boldsymbol{c} = \sigma^2 \sum_{i=1}^{n} c_i^2 = \tag{1.10}$$

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

小二乘估计量的方差

推导$\hat{\beta}_0$的方差表达式，用向量形式和Lemma (1.1)就十分方便，由$\{c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, i = 1, \ldots, n\}$

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n}\mathbf{1}'\boldsymbol{y} - \bar{x}\boldsymbol{c}'\boldsymbol{y}, \\
Var(\hat{\beta}_0) &= Var(\frac{1}{n}\mathbf{1}'\boldsymbol{y}) + Var(\bar{x}\boldsymbol{c}'\boldsymbol{y}) - 2Cov(\frac{1}{n}\mathbf{1}'\boldsymbol{y}, \bar{x}\boldsymbol{c}'\boldsymbol{y}) \\
&= \\
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + 0 = ? \qquad (1.11)
\end{aligned}$$

小二乘估计量的协方差(课堂练习十分钟)

提示，用到Lemma (1.1)以及向量表达

1. 同学们推导$Cov(\hat{\beta}_0, \hat{\beta}_1)$.

2. 同学们推导$Var(\hat{y}_*) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_*) =?$, x_* 是一个固定值。

一、t检验

得到经验归回方程$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$后，还需要进行检验，而检验就需要$\varepsilon_i \sim N(0, \sigma^2)$。一元回归分析中，要检验解释变量$x$是否对响应变量$y$的影响显著，就是做如下t检验

$$H_0 : \beta_1 = 0 \quad v.s. \quad H_1 : \beta_1 \neq 0$$

由(1.10)得，$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$, 于是在原假设$H_0$成立时，有

$$\hat{\beta}_1 \sim N(0, \sigma^2/S_{xx}), \tag{1.12}$$

说明此时，$\hat{\beta}_1$在零附近波动，又由于σ^2未知，因此可以构造t统计量

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}$$

t检验

题目：若已知$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 是σ^2的无偏估计，$(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$，且与$\hat{\beta}_1$独立，请同学们推导为什么$t \sim t_{n-2}$.

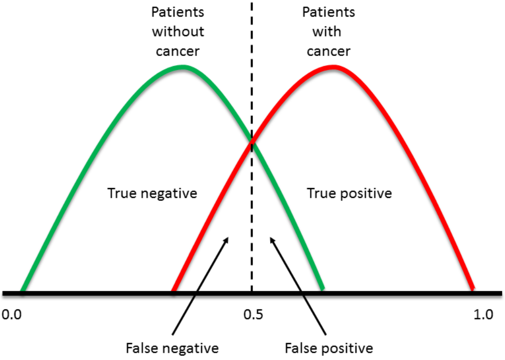
做t检验，有两种办法

1. 比较统计量t与其分布的分位数$t_{\alpha/2}$，即临界值。
2. 计算P-value，和显著性水平α比较。

提问，α的定义，犯第___类错误的概率。

Type I and Type II Errors

- 1-specificity=False Positive Rate; specificity= True Negative Rate
- sensitivity=True Positive Rate; 1-sensitivity=False Negative Rate
- False Positive Error is called Type I Error
- False Negative Error is called Type II Error



F检验

根据平方和分解，从回归效果来检验一元回归方程的显著性

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= SSR + SSE$$

请同学们补全推导过程。

SST就是反映y自身的波动，SSR就是由自变量x引起的y的波动（被解释的波动），SSE就是不能被x解释的波动。构造F检验统计量

$$F = \frac{SSR/1}{SSE/(n-2)} \sim \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2}$$

.

相关系数的显著性检验

对一元线性回归方程而言，可用$\{x_i, y_i, i = 1, \ldots, n\}$的相关系数来检验回归方程的显著性

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

r称为x, y的简单线性相关系数，简称相关系数，它表示x, y线性关系的密切程度，取值范围$|r| \leq 1$。

相关系数的直观意义图

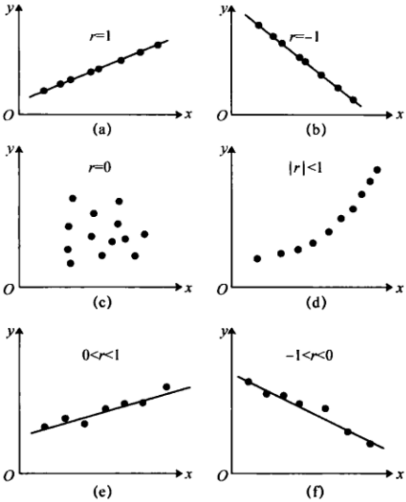


Figure 2: 若干情况的相关系数

相关系数的显著性检验

根据$\hat{\beta}_1$的表达式(1.8)得，(同学们推导一下)

$$r = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$$

一元线性回归中，r与$\hat{\beta}_1$的符号相同。

注意几点

1. 相关系数与n
2. 相关系数的检验：(I).临界值方法（查表）；(II).构造统计量，求P-value方法，因为$t = r\sqrt{n-2)}/\sqrt{1-r^2}$服从$t_{n-2}$分布。
3. 样本相关系数r与总体相关系数ρ。
   - I 例子，$r_A = 0.8$, 显著检验没通过；$r_B = 0.1$, 显著检验通过
   - II 例子，苏丹红与患癌率的$r_A = 0.2$; 保健品与健康长寿的$r_B = 0.2$

思考：如何解释？

三种检验的关系

对于一元线性回归模型而言，三种检验，t检验，F检验，相关系数检验是完全等价的。也就是说，只要一个检验拒绝H_0,其余也会拒绝；只要一个接受H_0, 其余也会接受。

1. $t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ 与$t=\frac{\hat{\beta}_1\sqrt{S_{xx}}}{\hat{\sigma}}$ 等价的
2. t_{n-2}与$F_{1,n-2}$也是等价的

同学们五分钟推导第一个，第二个略。

决定系数

决定系数(或判定系数，或确定系数)的定义如下

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

由关系式$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$ 可知，对一元线性回归模型而言，r^2正好是相关系数的平方。

几点注意：

1. r^2与样本量的关系
2. r^2很大，也不代表一定是线性关系
3. r^2大小与显著性检验的关系

总之，$|r|$或者r^2的大小，只是基于样本，对x和y的关联程度，所做的一个度量；而这个度量靠不靠谱，就是检验要做的事情。举一个例子，就像我们随机找一个人，让他说出自己的总资产（r^2就类似于他所说的资产），而他说的话可不可信，我们可以用测谎仪来判断（显著检验就如同测谎仪）。

残差的概念和残差图

前面的显著性检验是由假设条件的，$\varepsilon_i \sim N(0, \sigma^2)$。因此，我们必须诊断残差是否满足假设条件，才能放心运用回归分析。$e_i = y_i - \hat{y}_i$实际上就是ε_i的估计值。

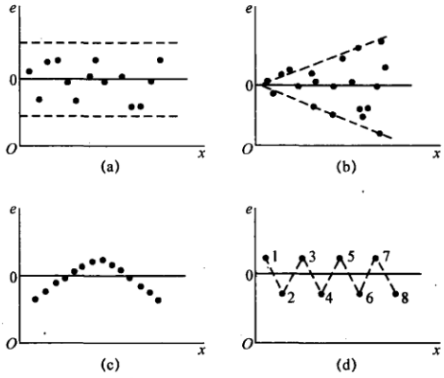


Figure 3: 残差图

几种常见残差图的解释

a. 回归模型基本满足假设条件
b. 异方差的出现，因此不满足σ^2为常数的假设。
c. 非线性关系的出现，因此不满足线性模型假设。
d. 周期性的出现，y的自相关性存在，因此不满足独立假设。

残差的性质

1. 性质一，$E(e_i) = 0$
2. 性质二，$Var(e_i) = (1 - h_{ii})\sigma^2$。同学们可以由$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$推导。这里，$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$称之为杠杆值，$0 < h_{ii} < 1$。解释$h_{ii}$大小与$(x_i - \bar{x})$的关系。
3. 性质三，残差满足约束条件，零均值$\mathbf{1}'\boldsymbol{e} = 0$, n维空间里与解释变量构成的向量正交，$\boldsymbol{x}'\boldsymbol{e} = 0$

残差的改进

由于残差的方差不相等，用e_i来判断模型假设会带来一定的麻烦，因此学者提出标准化(Standardized)残差和学生化(Studentized，William Sealy Gosset，Biometrika, 1908)残差。

1. $ZRE_i = \frac{e_i}{\hat{\sigma}}$
2. $SRE_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$

学生化残差进一步解决了方差不等的问题，因而更适合寻找异常值($SRE_i > 3$)。此外，在R程序里面

$$my\_mod < - \; lm(y \sim x1 + x2 + x3)$$
$$plot(my\_mod)$$

就可以将所有的残差图画出来用作诊断。

$\hat{\beta}$的区间估计

在最小二乘估计的性质一节里，我们推导出了$(\beta_0, \beta_1)^\top$ 各自的分布，由此可以构造出置信水平为$1-\alpha$的置信区间。这样的置信区间有三个特征

1. 区间是以$\{\hat{\beta}_j, j=0,1\}$为中心，以概率$1-\alpha$包含参数$\beta_j, j=0,1$
2. 区间长度越短，说明$\hat{\beta}_j, j=0,1$ 对$\beta_j, j=0,1$的估计精度越高
3. 区间长度越长，说明$\hat{\beta}_j, j=0,1$ 对$\beta_j, j=0,1$的估计精度越低

β_1的区间估计

实际中主要关注$\hat{\beta}_1$的估计精度，根据第三节推导的性质，式(1.9)以及(1.10),

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx}),$$

其中σ^2未知，可由$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2$ 构造t分布如下

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{S_{xx}}}{\hat{\sigma}} \sim t(n-2),$$

因而可根据t_{n-2}的分位数$t_{\alpha/2}(n-2)$建立

$$P\left(\left|\frac{(\hat{\beta}_1 - \beta_1)\sqrt{S_{xx}}}{\hat{\sigma}}\right| < t_{\alpha/2}(n-2)\right) = 1 - \alpha,$$

得到β_1的置信水平为$1-\alpha$的置信区间

$$\left(\hat{\beta}_1 - t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}},\quad \hat{\beta}_1 + t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right) \tag{1.13}$$

β_0的区间估计

请同学们根据(1.11)的结果

$$\hat{\beta}_0 \sim N\Big(\beta_0, \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}}\Big),$$

课堂练习：构造β_0的置信水平为$1-\alpha$的置信区间

$$\Big(\hat{\beta}_0 - \qquad\qquad\qquad , \quad \hat{\beta}_0 + \qquad\qquad\qquad \Big) \tag{1.14}$$

单值预测

当建立回归方程$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$之后，对解释变量的一个新观测值$x = x_*$，其单值预测为

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*,$$

即为$y_* = \beta_0 + \beta_1 x_* + \varepsilon_*$的单值预测。有两点注意，

1. y_*是一个随机变量，根据无偏估计的定义，$\hat{y}_*$不能称之为y_*的无偏估计
2. 而由
$$E(\hat{y}_*) = E(y_*) = \beta_0 + \beta_1 x_*, \tag{1.15}$$
他们是同均值的，因此可以说$\hat{y}_*$是y_*关于x_*条件均值的无偏估计

区间预测

单值预测只能给出期望值，不能确定预测精度，因此在给定显著性水平α下，找一个区间(T_1, T_2), 使得对某特定的x_*, y_*以概率$1-\alpha$落在区间(T_1, T_2)之内，即

$$P(T_1 < y_* < T_2) = 1-\alpha$$

分两种情况

1. 一种是对$y_* = \beta_0 + \beta_1 x_* + \varepsilon_*$进行预测，即对响应变量的新值进行预测，所以要考虑$\varepsilon_*$进去
2. 一种是对$E(y_*) = \beta_0 + \beta_1 x_*$进行预测，即对响应变量新值的期望进行预测

新值的区间预测

由$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$, 以及$Var(\hat{\beta}_0), Var(\hat{\beta}_1), Cov(\hat{\beta}_0, \hat{\beta}_1)$的结论可得，

$$\hat{y}_* \sim N\Big(\beta_0 + \beta_1 x_*, (\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}})\sigma^2\Big)$$

记$h_{00} = \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}$ 为新值x_*的杠杆值，则有$\hat{y}_* \sim N(\beta_0 + \beta_1 x_*, h_{00}\sigma^2)$。
观察，$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$, $\hat{\beta}_0$与$\hat{\beta}_1$ 都是$\boldsymbol{y} = (y_1, \ldots, y_n)^\top$的线性组合，新值$y_*$可以看成是第$n+1$个，与先前的$n$个观测值是独立的，因此

$$Var(y_* - \hat{y}_*) = Var(y_*) + Var(\hat{y}_*) - 2Cov(y_*, \hat{y}_*) = \sigma^2 + h_{00}\sigma^2.$$

响应变量新值的区间预测

且由式子(1.15), $E(y_* - \hat{y}_*) = 0$,

$$y_* - \hat{y}_* \sim N(0, (1 + h_{00})\sigma^2)$$

进一步，可得统计量

$$t = \frac{y_* - \hat{y}_*}{\sqrt{1 + h_{00}}\hat{\sigma}} \sim t(n-2)$$

由$P\left(\left|\frac{y_* - \hat{y}_*}{\sqrt{1+h_{00}}\hat{\sigma}}\right| < t_{\alpha/2}(n-2)\right) = 1 - \alpha$, 可得$y_*$的置信水平为$1 - \alpha$的置信区间为：

$$\left(\hat{y}_* - t_{\alpha/2}(n-2)\sqrt{1 + h_{00}}\hat{\sigma}, \quad \hat{y}_* + t_{\alpha/2}(n-2)\sqrt{1 + h_{00}}\hat{\sigma}\right) \tag{1.16}$$

响应变量新值的区间预测

影响(1.16)中y_*预测精度的几个因素如下：

1. 样本量n越大，区间越短
2. $S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=(n-1)S_x^2$, 基本上是随着$n$增大而增大，区间也会更短
3. x_*靠近$\bar{x}$, 区间越短

当样本量n很大，$|x_*-\bar{x}|$较小时，h_{00}接近零，这时可以近似构造y_*的95%置信区间：

$$\hat{y}_*\pm 2\hat{\sigma}. \tag{1.17}$$

响应变量新值期望的区间预测

式(1.16)提供的是响应变量新值的置信区间，往往人们还关心另一种情况，即响应变量期望（均值）的区间估计，也即是常数$E(y_*)$的区间估计。由，

$$\hat{y}_* - E(y_*) \sim N(0, h_{00}\sigma^2)$$

进而的置信水平为$1-\alpha$的置信区间

$$\left(\hat{y}_* - t_{\alpha/2}(n-2)\sqrt{h_{00}}\hat{\sigma},\quad \hat{y}_* + t_{\alpha/2}(n-2)\sqrt{h_{00}}\hat{\sigma}\right) \tag{1.18}$$

请同学们对比式(1.16)和(1.18)，他们有什么区别，产生区别的原因是什么？

控制问题

控制问题可以看作预测问题的逆问题，即如何控制解释变量x的值从而以$1-\alpha$的概率保证响应变量y的值控制在$T_1 < y < T_2$中，即

$$P(T_1 < y < T_2) = 1 - \alpha,$$

通常用近似的预测区间来确定x。一般设定$\alpha = 0.05$，根据(1.17), 构造不等式组

$$\begin{cases} \hat{y}(x) - 2\hat{\sigma} > T_1 \\ \hat{y}(x) + 2\hat{\sigma} < T_2 \end{cases} \tag{1.19}$$

将$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$代入，得到, 当$\hat{\beta}_1 > 0$ 时

$$\frac{T_1 + 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \frac{T_2 - 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} \tag{1.20}$$

当$\hat{\beta}_1 < 0$ 时

$$\frac{T_2 - 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \frac{T_1 + 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} \tag{1.21}$$

总结

1. 最小二乘估计原理和性质是重点，学会计算，并掌握推导、证明
2. 显著性检验和残差分析是难点，理解原理，会运用检验解决实际问题，掌握分析方法
3. 区间估计、预测和控制会纳入考点，结合重点、难点内容进行考察
4. 自学仔细阅读教材—2.8 本章小结与评注，对完成小课题大有帮助