简单线性回归

最小二乘法估计的性质

$y=\beta_0+\beta_1 x+\epsilon\ ,\quad E(\epsilon|x)=0\ ,\quad Var(\epsilon|x)=\sigma^2$

假设(x_i,y_i)是来自(x,y)的随机样本，则模型也可以表示为：

$y_i=\beta_0+\beta_1 x_i+\epsilon_i\ ,E(\epsilon_i)=0\ ,\quad Var(\epsilon_i)=\sigma^2$

上面介绍最小化残差平方和$\alpha=\sum_{i=1}^n(y_i-\hat\beta_0-\hat\beta_1 x)^2$可得OLS估计为

$\hat\beta_1=\frac{\sum_{i=1}^n y_i x_i-n\bar x\bar y}{\sum_{i=1}^n x_i^2-n\bar x^2}=\frac{\sum_{i=1}^n(x_i-\bar x)(y_i-\bar y)}{\sum_{i=1}^n(x_i-\bar x)^2}=\frac{\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar x)(y_i-\bar y)}{\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar x)^2}=\frac{S_{xy}}{S_{xx}}$

$\hat\beta_0=\bar y-\hat\beta_1\bar x$

下面我们将通过求偏导数得到最小化平方和的参数

$$\begin{cases}\frac{\partial\alpha}{\partial\hat\beta_0}\Big|_{\hat\beta_0,\hat\beta_1}=0\\ \frac{\partial\alpha}{\partial\hat\beta_1}\Big|_{\hat\beta_0,\hat\beta_1}=0\end{cases}\Longrightarrow\begin{cases}\sum_{i=1}^n(y_i-\hat\beta_0-\hat\beta_1 x_i)\Big|_{\hat\beta_0,\hat\beta_1}=0\\ \sum_{i=1}^n(y_i-\hat\beta_0-\hat\beta_1 x_i)(x_i)\Big|_{\hat\beta_0,\hat\beta_1}=0\end{cases}\Longrightarrow\begin{cases}\sum_{i=1}^n e_i=0\\ \sum_{i=1}^n e_i x_i=0\end{cases}\quad(*)$$

- 性质：
  - 1️⃣ $\sum_{i=1}^n e_i=0$
  - 2️⃣ $(\bar x,\bar y)$在样本回归线上
  - 3️⃣ e_i和x_i样本相关系数为0，e_i和x_i不相关
    - $\sum e_i x_i=\sum(e_i-\bar e)x_i=\sum(e_i-\bar e)(x_i-\bar x)=\sum(e_i-\bar e)x_i-\bar x\sum(e_i-\bar e)=$
  - 4️⃣ e_i和$\hat y_i$不相关
    - $\sum e_i\hat y_i=\sum e_i(\hat\beta_0+\hat\beta_1 x_i)=\hat\beta_0\sum e_i+\hat\beta_1\sum e_i x_i=0=\sum(e_i-\bar e)(y_i-\bar y)$

下面我们将会研究OLS估计量的统计性质，具体地，我们将会证明

- OLS估计为线性估计，即$\hat\beta_1$可以表示为$y_1,\cdots,y_n$线性组合形式

$$\begin{aligned}\hat\beta_1&=\frac{S_{xy}}{S_xx}\\&=\frac{\sum(s_i-\bar x)(y_i-\bar y)}{\sum(x_i-\bar x)^2}\\&=\sum\frac{x_i-\bar x}{S_{xx}}y_i-\sum\frac{x_i-\bar x}{S_{xx}}\bar y\\&=\sum\frac{x_i-\bar x}{S_{xx}}y_i\\&=\sum c_i y_i\\&=\sum c_i(\beta_0+\beta_1 x_i+\epsilon_i)\\&=\beta_0\sum c_i+\beta_1\sum c_i x_i+\sum c_i\epsilon_i\\&=\beta_1+\sum c_i\epsilon_i\end{aligned}$$

$$c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

$$E(\hat{\beta}_1) = \beta_1 + \sum c_i E(\epsilon_i) = \beta_1$$

$$bias(\hat{\beta}_1) = E(\hat{\beta}_1) - \beta_1 = 0 \quad \hat{\beta}_1\text{为}\beta_1\text{的线性无偏估计}$$

$$\text{同理，我们可以得到}$$

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\
&= \frac{1}{n}\sum y_i - \bar{x}\sum c_i y_i \\
&= \sum[\frac{1}{n} - \bar{x}c_i]y_i \\
&= \sum d_i y_i \\
&= \beta_0 \sum d_i + \beta_1 \sum d_i x_i + \sum d_i \epsilon_i \\
&= \beta_0 + \sum d_i \epsilon_i
\end{aligned}$$

$$E(\hat{\beta}_0) = E(\beta_0) = \beta_0$$

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var(\beta_1 + \sum c_i \epsilon_i) \\
&= Var(\sum c_i \epsilon_i) \\
&= \sum c_i^2 \sigma^2 \\
&= \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{S_{xx}^2} \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

$$\begin{aligned}
Var(\hat{\beta}_0) &= \sum d_i^2 \sigma^2 \\
&= \sum(\frac{1}{n} - \bar{x}\frac{x_i - \bar{x}}{S_{xx}})^2 \sigma^2 \\
&= (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2
\end{aligned}$$

书上的证明方法

$$\begin{aligned}
Var(\hat{\beta}_0) &= Var(\bar{y} - \bar{x}\hat{\beta}_1) \\
&= Var(\bar{y}) + Var(\bar{x}\hat{\beta}_1) - 2Cov(\bar{y}, \bar{x}\hat{\beta}_1) \\
&= \frac{1}{n^2}\sum Var(\epsilon_i) + \bar{x}^2 Var(\hat{\beta}_1) \\
&= (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2
\end{aligned}$$

随机误差方差σ^2 的估计

SLR$(simple\ linear\ regression)$

$$y_i = \beta_0 + \beta_1 x + \epsilon_i\ , \quad \epsilon_i \sim (0, \sigma^2)$$

通过上面的计算，我们可以得到以下的参数估计的均方误差

$$MSE(\hat{\beta}_1) = Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$MSE(\hat{\beta}_0) = Var(\hat{\beta}_0) = (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2$$

从上面的式子中，我们可以发现，在估计参数的均方误差的过程，实际上我们只需要进行估计σ即可，接下来，我们将会对估计σ展开工作。

首先估计σ^2有以下的式子，但是在本例子中，我们只有样本数据，因此我们需要通过样本数据残差e_i来对ϵ_i进行一个估计

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\sum e_i^2 \\
&= \frac{1}{n-2}\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\end{aligned}
$$

此处$n-2$表示的是自由度，在$(*)$中有两个限制条件

$$
\begin{aligned}
&\Longrightarrow E(\hat{\sigma}^2) = \sigma^2 \\
&\Longrightarrow E(SS_{res}) = E(\sum e_i^2) = (n-2)E(\hat{\sigma}^2) = (n-2)\sigma^2 \\
&\Longrightarrow \frac{SS_{res}}{n-2}\text{为}\sigma^2\text{的无偏估计} \\
&\Longrightarrow \hat{\sigma}^2 = \frac{SS_{res}}{n-2} = MS_{res}
\end{aligned}
$$

下面引入一个概念回归标准误差

$$
s.e.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \sqrt{\frac{SS_{res}}{S_{xx}(n-2)}}
$$

$$
s.e.(\hat{\beta}_0) = \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\hat{\sigma} = \sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\frac{SS_{res}}{n-2}}
$$

火箭推进器数据代码

```r
# 首先先进行读取数据
da2_1 = readxl::read_xls("./DataSets4e/Chapter 2/Examples/data-ex-2-1 (Rocket Prop).xls")
colnames(da2_1) <- c("ID","y","x")
da2_1
# 下面使用fit函数对其进行拟合
fit <- lm(y~x,data = da2_1)
res <- resid(fit) # 抽取残差值序列
SS_res <- sum(res^2)
# 下面直接进行输出残差平方和
SS_res


# 下面我们先进行估计对应的sigma_hat
# 首先进行读取维度的信息
n <- dim(da2_1)[1]
df <- n-2
sigma_hat <- sqrt(SS_res/df)
# 下面进行输出sigma_hat
sigma_hat


# 接下来使用对sigma的参数估计对参数的标准误差进行估计
x = da2_1$x
Sxx = sum((x-mean(x))^2)
s.e.beta1 <- sigma_hat/sqrt(Sxx)
```

```
s.e.beta0 <- sigma_hat*sqrt((1/n + mean(x)^2/Sxx))
s.e.beta1
s.e.beta0

# 使用函数的形式对拟合的数据进行输出
summary(fit)
```

简单线性模型的另外一种表达形式

对于简单线性模型的原本的形式：

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

对于给定的数据，我们能够将上面的表达式进行变形

$$\begin{aligned} y_i &= \beta_0 + \beta_1(x_i - \bar{x}) + \beta_1\bar{x} + \epsilon_i \\ &= (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + \epsilon_i \\ &= \beta_0' + \beta_1(x_i - \bar{x}) + \epsilon_i \\ &= \beta_0' + \beta_1 x_i' + \epsilon_i \end{aligned}$$

将初始的形式转换为下面的标准的形式的时候，对于斜率的参数估计不会改变，但是对于截距的估计是会改变的

$$\text{标准形式下的}\beta_0\text{和}\beta_1\text{的参数估计和原始形式下的}\beta_0\text{和}\beta_1$$
$$\hat{\beta}_1 = \frac{S_{x'y}}{S_{x'x'}} = \frac{\sum(x' - \bar{x}')(y - \bar{y})}{\sum(x' - \bar{x}')^2} \quad \hat{\beta}_0' = \bar{y} - \hat{\beta}_1\bar{x}' = \bar{y}$$
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- 将简单的线性模型进行标准化的特点如下
  - 此时截距的估计和斜率的估计是不相关的，即$Cov(\hat{\beta}_0', \hat{\beta}_1) = 0$，对于第二个式子，我们可以发现
  - 如果我们用该模型进行预测，则有$\hat{y} = \bar{y} + \hat{\beta}_1 \cdot (x - \bar{x})$，提醒我们回归模型的有效范围在其均值点附近