第2次作业

姓名：韩璐瑶　　学号：1847405023

1　证明无β_0时，$SST = SSR + SSE$成立。

证

因为

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \end{aligned}$$

其中

$$\begin{aligned} &\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ =& \sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)(\hat{\beta}_1 x_i - \bar{y}) \\ =& \hat{\beta}_1 \sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)x_i - \bar{y}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i) \end{aligned}$$

由于一元线性回归方程是在离差平方和达到最小时建立，故根据最小二乘法原理，有

$$\frac{dQ}{d\beta_1}\Big|_{\beta_1 = \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)x_i = 0$$

且

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i = n\bar{y} - n\hat{\beta}_1\bar{x} = 0$$

所以

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$$

因此

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) \\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\ &= SSR+SSE \end{aligned}$$

2 判定r^2是否一定大于零。

答

由关系式

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

知$r^2\geq 0$，又根据关系式

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

可知，当$\hat{\beta}_1=0$或$\sum_{i=1}^{n}(x_i-\bar{x})^2=0$时，$r^2=0$。而由于$\sum_{i=1}^{n}(x_i-\bar{x})^2\neq 0$（不考虑只有一个观测值），因此仅当$\hat{\beta}_1=0$时，$r^2=0$，此时，$\hat{y}_i=\hat{\beta}_0$，自变量不影响因变量，回归效果极差。

所以，只要$\hat{\beta}_1\neq 0$，r^2大于零一定成立。

3 验证三种检验的关系：

(1) $$t=\frac{\hat{\beta}_1\sqrt{L_{xx}}}{\hat{\sigma}}=\frac{\sqrt{n-2}r}{\sqrt{1-r^2}}$$

(2) $$F=\frac{SSR/1}{SSE/(n-2)}=\frac{\hat{\beta}_1^2L_{xx}}{\hat{\sigma}^2}=t^2$$

证

（1） 因为

$$\begin{cases} \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{SSE}{n-2} \\ \hat{\beta}_1^2 L_{xx} = SSR \end{cases}$$

所以有

$$t^2 = \frac{\hat{\beta}_1^2 L_{xx}}{\hat{\sigma}^2} = \frac{(n-2)SSR}{SSE}$$

又因为

$$r^2 = \frac{SSR}{SST}$$

所以

$$\frac{(n-2)r^2}{1-r^2} = \frac{(n-2)SSR}{SST-SSR} = \frac{(n-2)SSR}{SSE} = t^2$$

故（1）式成立。

（2） 由（1）知，有

$$t^2 = \frac{\hat{\beta}_1^2 L_{xx}}{\hat{\sigma}^2} = \frac{(n-2)SSR}{SSE} = \frac{SSR/1}{SSE/(n-2)}$$

故（2）式显然成立。

4 验证决定系数r^2与F值之间的关系 $r^2 = \frac{F}{F+n-2}$。该表达式说明r^2与F值是等价的，那么我们为什么要分别引入这两个统计量，而不是只使用其中的一个？

验证

$$\frac{F}{F+n-2} = \frac{(n-2)SSR/SSE}{(n-2)SSR/SSE+(n-2)SSE} = \frac{(n-2)SSR}{(n-2)SSR+(n-2)SSE} = \frac{SSR}{SST} = r^2$$

解释

F值与决定系数r^2均可以表示总离差平方和中回归平方和与残差平方和的占比的大小，但F值考虑了回归平方和与残差平方和的自由度，在样本量n较小时，可以较为准确的反映回归直线与样本观测值的拟合优度。而r^2在样本量n较大时，可以较为简单直接的反映拟合优度，两个各有利弊，因此分别引入两个统计量。

5　为了调查某广告对销售收入的影响，某商店记录了5月份的销售收入y（万元）和广告费用x（万元），调查数据如表1所示。

表 1: 销售收入与广告费用

月份	1	2	3	4	5
x	1	2	3	4	5
y	10	10	20	20	40

(1)　画散点图

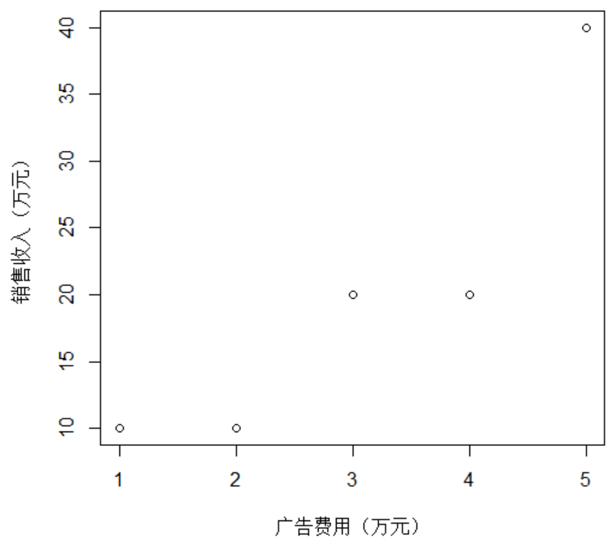

图 1: 散点图

(2)　x与y之间是否大致呈线性关系？

由图1可见，x与y之间大致呈线性关系。

(3)　用最小二乘估计求出回归方程。

设回归方程为$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$。

根据$\hat{\beta}_0$和$\hat{\beta}_1$的最小二乘计算公式

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = L_{xy}/L_{xx} \end{cases}$$

编程计算出$\hat{\beta}_0$和$\hat{\beta}_1$，分别为 -1 和 7。故回归方程为

$$\hat{y} = -1 + 7x$$

```
x <- c (1, 2, 3, 4, 5) ;
y <- c (10, 10, 20, 20, 40) ;
plot(x, y, xlab = "广告费用（万元）", ylab = "销售收入（万元）")
n <- 5;
L_xx <- sum( x ^ 2 )- n * mean( x )^2;
L_yy <- sum( y ^ 2 )- n * mean( y )^2;
L_xy <- sum( x * y )- n * mean( x ) * mean( y );
beta_1 <- L_xy / L_xx;                          #计算beta_估计值1
beta_0 <- mean( y ) - beta_1 * mean( x );       #计算beta_估计值0
```

Code of (1)(3)

(4) 求回归标准误差$\hat{\sigma}$。

由最大似然估计可得σ^2的估计值为

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

根据该公式编程计算得回归标准误差$\hat{\sigma}$为 6.0553。

(5) 给出$\hat{\beta}_0$与$\hat{\beta}_1$的置信度为 95%的区间估计。

由于$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$，所以检验统计量

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

服从自由度为 3 的t分布。因而

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

即得β_1的置信度为$1-\alpha$的置信区间为

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

根据上述公式（其中$\alpha = 0.05$）编写程序，计算得$\hat{\beta}_1$的置信度为95%的置信区间为(0.9061, 13.0940)。

由于$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{(\bar{x})^2}{L_{xx}}\right)\sigma^2\right)$，同理可计算$\hat{\beta}_0$的置信度为95%的置信区间为(−21.2112, 19.2112)。

(6) 计算x与y的决定系数。

根据关系式

$$r^2 = \frac{SSR}{SST} = \frac{L_{xy}^2}{L_{xx}L_{yy}}$$

编程计算得x与y的决定系数为0.8167。

```
y_hat <- -1 + 7 * x
sigma <- sqrt( 1/( n - 2 ) * sum( ( y - y_hat ) ^ 2 ) )
alpha <- 0.05
#计算beta_1 置信区间的端点值
tmp_1 <- qt(1 - alpha / 2, n - 2) * sigma / sqrt( L_xx );
a_1 <- beta_1 - tmp_1
b_1 <- beta_1 + tmp_1
#计算beta_0 置信区间的端点值
tmp_0 <- qt(1 - alpha / 2, n - 2) * sqrt(1/n + mean(x)^2 / L_xx) * sigma
a_0 <- beta_0 - tmp_0
b_0 <- beta_0 + tmp_0
#计算 x 与 y 的决定系数
r2 = L_xy ^ 2 / L_xx / L_yy
```

Code of (4)(5)(6)

(7) 对回归方程做方差分析。

- 原假设 H_0 :回归方程不显著，备择假设 H_1 :回归方程显著。
- 计算方差分析表

表 2: 方差分析表

方差来源	自由度	平方和	均方	F值	P值
回归	1	490	490	13.3636	0.0354
残差	3	110	36.6667		
总和	4	600			

- 当取 $\alpha = 0.05$ 时，$P < \alpha$，故拒绝原假设，认为回归方程显著，认为x与y有显著的线性关系。

(8) 做回归系数$\hat{\beta}_1$的显著性检验。

- 原假设 $H_0 : \beta_1 = 0$，备择假设 $H_1 : \beta_1 \neq 0$。

- 构造F检验统计量

$$F = \frac{SSR/1}{SSE(n-2)}$$

正态假设下，原假设成立时，F服从自由度为$(1, n-2)$的F分布。
- 当F值大于临界值$F_\alpha(1, n-2)$时，拒绝H_0。
- 当取$\alpha = 0.05$时，$P < \alpha$，故拒绝原假设，认为x对y有显著的影响。

(9) 做相关系数的显著性检验。

根据关系式

$$r = \frac{L_{xx}}{\sqrt{L_{xx}L_{yy}}}$$

计算得r为0.9037。但由于n较小，仅凭相关系数较大就说明变量x与y之间有密切得线性关系颇为草率。

由于$n = 5$，课本附录相关系数的检验表中对应$\alpha = 5\%$，$n - 2 = 3$对应的值为0.878，$\alpha = 1\%$，$n - 2 = 3$对应的值为0.959，而$0.878 < r = 0.9037 < 0.959$，因此说明$x$与$y$有显著的线性关系。

(10) 对回归方程作残差图并做相应的分析。

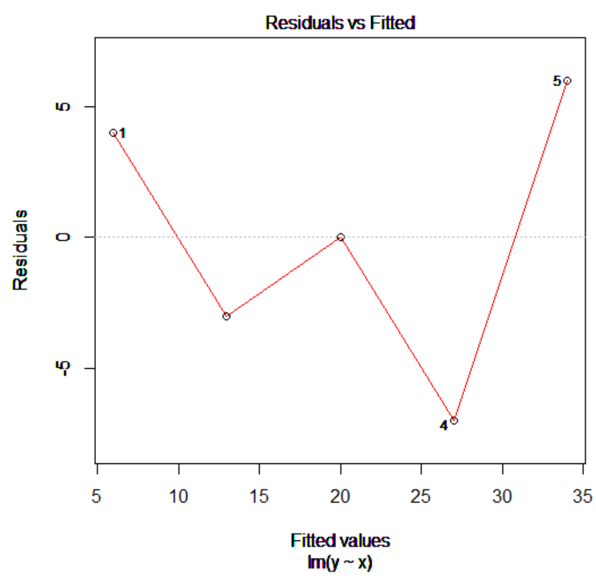

图 2: 残差图

从图上我们可以看出，残差在$e = 0$附近变化，且在变化幅度不大的一个区域。但是，由于观测个数较小，所以可能会带来较大的误差。

(11) 求当广告费用为4.2万元时，销售收入将达到多少，并给出置信度为95%的置信区间。

根据回归方程 $\hat{y}_0 = -1 + 7x_0$ 可知，当 $x_0 = 4.2$ 时，计算得 $\hat{y}_0 = 28.4$。
由于

$$\hat{y}_0 \sim N\Big(-1 + 7x_0, \Big(\frac{1}{n} + \frac{(x_0 - x)^2)}{L_{xx}}\Big)\sigma^2\Big)$$

所以统计量

$$t = \frac{y_0 - \hat{y}_0}{\sqrt{1 + h_{00}}\hat{\sigma}} \sim t(n-2)$$

其中

$$h_{00} = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}$$

可得

$$P\Bigg(\Bigg|\frac{y_0 - \hat{y}_0}{\sqrt{1 + h_{00}}\hat{\sigma}}\Bigg| < t_{\alpha/2}(n-2)\Bigg) = 1 - \alpha$$

所以 y_0 的置信度为 $1 - \alpha$ 的置信区间为

$$\Big(y_0 - t_{\alpha/2}(n-2)\sqrt{1 + h_{00}}\hat{\sigma}, y_0 + t_{\alpha/2}(n-2)\sqrt{1 + h_{00}}\hat{\sigma}\Big)$$

根据上述公式（其中$\alpha = 0.05$）编写程序，计算得 y_0 的置信度为95%的置信区间为(6.0593, 50.7407)。

```
#画残差图
fit <- lm (y ~ x, data = e)
plot(fit, which = 1)
#计算y_0
x_0 <- 4.2;
y_0 <- -1 + 7 * x_0;
#计算y_0 置信区间的端点值
h_00 <- 1 / n + ( x_0 - mean( x ) ) ^ 2 / L_xx;
tmp_y <- qt (1 - alpha / 2, n - 2) * sqrt ( 1 + h_00) * sigma;
a_y <- y_0 - tmp_y;
b_0 <- y_0 + tmp_y;
```

Code of (10)(11)