第四次作业

1907402030 熊雄

2021 年 10 月 23 日

题目 1. **(lec3.pdf P26)**

为了确保 $\hat{\boldsymbol{\beta}}_c$ 的确为约束条件下的 $\boldsymbol{\beta}$ 的最小二乘估计，我们还需要证明下面两点：

1. $\boldsymbol{A}\hat{\boldsymbol{\beta}}_c = \boldsymbol{b}$.
2. 对一切满足条件的 $\boldsymbol{A\beta} = \boldsymbol{b}$ 的 $\boldsymbol{\beta}$, 都有 $\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2 \geq \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_c\|^2$.

解答.

1. **Proof.**

$$\begin{aligned}
\boldsymbol{A}\hat{\boldsymbol{\beta}}_c &= \boldsymbol{A}\left[\hat{\boldsymbol{\beta}} - \left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{A^T}\left[\boldsymbol{A}\left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{A^T}\right]^{-1}\left(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b}\right)\right] \\
&= \boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{A}\left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{A^T}\left[\boldsymbol{A}\left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{A^T}\right]^{-1}\left(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b}\right) \\
&= \boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{A}\hat{\boldsymbol{\beta}} + \boldsymbol{b} \\
&= \boldsymbol{b}.
\end{aligned}$$

2. **Proof.**

由于

$$\hat{\boldsymbol{\beta}}_c = \left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{X^T Y} - \left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{A^T}\hat{\boldsymbol{\lambda}}_c = \hat{\boldsymbol{\beta}} - \left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{A^T}\hat{\boldsymbol{\lambda}}_c$$

故可以导出下述关系：

$$\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_c\right)^{\boldsymbol{T}}\boldsymbol{X^T X}\left(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}\right) = \hat{\boldsymbol{\lambda}}_c^{\boldsymbol{T}}\boldsymbol{A}\left(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}\right) = \hat{\boldsymbol{\lambda}}_c^{\boldsymbol{T}}\left(\boldsymbol{A}\hat{\boldsymbol{\beta}}_c - \boldsymbol{A\beta}\right) = 0$$

于是, 对不等式左边作如下分解:

$$
\begin{aligned}
\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 &= \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\\
&= \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c+\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta}\right)^T\boldsymbol{X}^T\boldsymbol{X}\\
&\quad \left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c+\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta}\right)\\
&= \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c\right)+\\
&\quad \left(\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta}\right)\\
&= \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c\right)\right\|^2+\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta}\right)\right\|^2.
\end{aligned}
$$

因此, 我们可以得到对一切满足条件的 $\boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b}$ 的 $\boldsymbol{\beta}$, 总有

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \geq \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c\right)\right\|^2,$$

且等号成立当且仅当第三项 $\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta}\right)\right\|^2=0$, 也就是 $\boldsymbol{X}\hat{\boldsymbol{\beta}}_c=\boldsymbol{X}\boldsymbol{\beta}$. 于是用 $\boldsymbol{X}\hat{\boldsymbol{\beta}}_c$ 代替 $\boldsymbol{X}\boldsymbol{\beta}$, 等式成立, 即:

$$\left\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c\right\|^2=\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c\right)\right\|^2.$$

从而显然可以得到不等式:

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \geq \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c\|^2.$$

证毕.

题目 2. (课本 P84 d3.11)

研究货运总量 y（万吨）与工业总产值 x_1（亿元)、农业总产值 x_2（亿元)、居民非商品支出 x_3（亿元）的关系。

1. 计算出 y, x_1, x_2, x_3 的相关系数矩阵。
2. 求出 y 与 x_1, x_2, x_3 的三元线性回归方程。
3. 对所求的方程作拟合优度检验。

4. 对回归方程作显著性检验。

5. 对每一个回归系数作显著性检验。

6. 如果有的回归系数没有通过显著性检验，将其剔除，重新建立回归方程，并作回归方程的显著性检验和回归系数的显著性检验。

7. 求出每一个回归系数的置信水平为 95% 的置信区间。

8. 求标准化回归方程。

9. 求当 $x_{01}=75, x_{02}=42, x_{03}=3.1$ 时的 $\hat{y}_0$，并请给出置信水平为 95% 的置信区间。

10. 结合回归方程对问题做一些基本分析。

编号	货运总量 y (万吨)	工业总产值 x_1 (亿元)	农业总产值 x_2 (亿元)	居民非商品支出 x_3 (亿元)
1	160	70	35	1
2	260	75	40	2.4
3	210	65	40	2
4	265	74	42	3
5	240	72	38	1.2
6	220	68	45	1.5
7	275	78	42	4
8	160	66	36	2
9	275	70	44	3.2
10	250	65	42	3

解答. 先利用 **R** 建立数据集, 输入以下代码:

```r
n <- 10
x1 <- c(70, 75, 65, 74, 72, 68, 78, 66, 70, 65)
x2 <- c(35, 40, 40, 42, 38, 45, 42, 36, 44, 42)
x3 <- c(1, 2.4, 2, 3, 1.2, 1.5, 4, 2, 3.2, 3)
y <- c(160, 260, 210, 265, 240, 220, 275, 160, 275, 250)
dat <- do.call(cbind,list(y,x1,x2,x3))
```

1. **Solve.**

   输入以下代码用于计算相关系数矩阵：

```
cor(dat,method = "spearman")
```

   得到输出结果如下，即 y, x_1, x_2, x_3 的相关系数矩阵：

```
          [,1]       [,2]       [,3]      [,4]
[1,] 1.0000000 0.60122699 0.64399063 0.8435583
[2,] 0.6012270 1.00000000 0.04024941 0.2944785
[3,] 0.6439906 0.04024941 1.00000000 0.5789723
[4,] 0.8435583 0.29447853 0.57897234 1.0000000
```

2. **Solve.**

   继续输入以下代码：

```
lm(y ~ x1+ x2+ x3)
```

   输出后可以得到三元线性回归方程为:

$$y = 3.754x_1 + 7.101x_2 + 12.447x_3 - 348.280.$$

3. **Solve.**

   继续输入以下代码：

```
print(summary(lm(y ~ x1+ x2+ x3)))
```

   得到输出如下：

```
Residuals:
   Min      1Q  Median     3Q     Max
-25.198 -17.035   2.627  11.677  33.225

 Coefficients :
   Estimate Std. Error t value Pr(>|t|)
(Intercept) -348.280   176.459  -1.974   0.0959 .
x1             3.754     1.933   1.942   0.1002
x2             7.101     2.880   2.465   0.0488 *
x3            12.447    10.569   1.178   0.2835
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
    0.1 ' ' 1

Residual standard error: 23.44 on 6 degrees of freedom
Multiple R-squared:  0.8055,  Adjusted R-squared:  0.7083
F-statistic: 8.283 on 3 and 6 DF,  p-value: 0.01487
```

Multiple R-squared 和 Adjusted R-squared 这两个值我们常称之为"拟合优度"和"修正的拟合优度"，指的是回归方程对样本的拟合程度。在这里我们可以看到修正的拟合优度为 0.7083，比较接近 1，说明拟合程度较好。

4. **Solve.**

由第 3 问的输出结果的最后一行可以看到，**F** 检验的 p 值为 $0.01487 < 0.05$，故在显著性水平 0.05 下，我们认为通过回归方程整体的显著性检验。

5. **Solve.**

由第 3 问的输出结果可以看出，x_1, x_2 的 p 值分别为 0.1002 和 0.0488，说明回归系数较显著。而 x_3 的 p 值为 $0.2835 > 0.05$，说明 x_3 的回归系数不显著，应该予以剔除。

6. **Solve.**

重新输入以下代码:

```
n <- 10
x1 <- c(70, 75, 65, 74, 72, 68, 78, 66, 70, 65)
x2 <- c(35, 40, 40, 42, 38, 45, 42, 36, 44, 42)
y <- c(160, 260, 210, 265, 240, 220, 275, 160, 275, 250)
dat1 <- do.call(cbind,list(y,x1,x2))
fit <- lm(y ~ x1+ x2)
print(summary(lm(y ~ x1+ x2)))
```

运行后可以得到

```
Call:
  lm(formula = y ~ x1 + x2)

```

```
Residuals:
    Min      1Q  Median      3Q     Max
-42.012 -10.656   4.358  11.984  28.927

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -459.624    153.058  -3.003  0.01986 *
x1             4.676      1.816   2.575  0.03676 *
x2             8.971      2.468   3.634  0.00835 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.08 on 7 degrees of freedom
Multiple R-squared:  0.7605,  Adjusted R-squared:  0.6921
F-statistic: 11.12 on 2 and 7 DF,  p-value: 0.006718
```

故重新建立的回归方程为:

$$y = 4.676x_1 + 8.971x_2 - 459.624.$$

由输出结果的最后一行可以看到, **F** 检验的 p 值为 $0.006718 < 0.05$, 故在显著性水平 0.05 下, 我们认为通过回归方程整体的显著性检验。由系数表可知, 各个回归系数的 p 值较小, 说明回归系数的显著性较高。

7. **Solve.**

```
confint(fit, level=.9)
```

提交后输出得到:

```
                    5 %       95 %
(Intercept) -749.603256 -169.64405
x1             1.234941    8.11632
x2             4.294268   13.64765
```

因此, x_1 的置信度为 95% 的置信区间为 $(1.234941, 8.11632)$, x_2 的置信度为 95% 的置信区间为 $(4.294268, 13.64765)$.

8. **Solve.**

重新输入以下代码，对样本进行标准化之后求回归方程:

```
n <- 10
x1 <- c(70, 75, 65, 74, 72, 68, 78, 66, 70, 65)
x2 <- c(35, 40, 40, 42, 38, 45, 42, 36, 44, 42)
y <- c(160, 260, 210, 265, 240, 220, 275, 160, 275, 250)
dat1 <- do.call(cbind,list(y,x1,x2))
dat2 <- scale(dat1,center = T, scale = T)
lm(dat2[,1] ~ dat2[,2]+dat2 [,3])
```

输出后可以得到标准化后的回归方程为:

$$y = 0.4792x_1 + 0.6765x_2 - 7.552 \times 10^{-16}.$$

9. **Solve.**

重新输入以下代码:

```
n <- 10
x1 <- c(70, 75, 65, 74, 72, 68, 78, 66, 70, 65)
x2 <- c(35, 40, 40, 42, 38, 45, 42, 36, 44, 42)
y <- c(160, 260, 210, 265, 240, 220, 275, 160, 275, 250)
dat1 <- do.call(cbind,list(y,x1,x2))
 fit  <- lm(y ~ x1+ x2)
new <- data.frame(x1 = 75, x2 = 42, x3 = 3.1)
lm.pred <- predict(fit,new,interval = "prediction",level =
        0.95)
lm.pred
```

输出后可以得到:

```
> lm.pred
 fit      lwr      upr
1 267.829 204.4355 331.2225
```

即令 $x_{01} = 75, x_{02} = 42, x_{03} = 3.1$ 时，得到 $\hat{y}_0 = 267.829$. 同时置信区间为 $(204.4355, 331.2225)$.

10. **Solve.**

根据前面的分析，虽然 R^2 的值较大，但并不能说明回归方程显著，我们还需要通过对回归方程以及其系数进行检验。当一个回归方程通过

显著性检验之后，也并不能说明这个方程中所以自变量都对因变量 y 有显著影响，还需对回归系数进行检验。