Linear Model Assignment 4

110024516 統研碩一邱繼賢

2021 年 11 月 14 日

Problem 1.
Construct the full model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

```r
data = read.table("wastes.txt", skip = 1)
names(data) = c("day", "x1", "x2", "x3", "x4", "x5", "y")
g = lm(y ~ x1+x2+x3+x4+x5, data = data)
```

a.

$$CI = \begin{cases} (\hat{\beta}_3 - t_{(0.025,14)}\ se_{\hat{\beta}_3}\ ,\ \hat{\beta}_3 + t_{(0.025,14)}\ se_{\hat{\beta}_3}) \\ (\hat{\beta}_5 - t_{(0.025,14)}\ se_{\hat{\beta}_5}\ ,\ \hat{\beta}_5 + t_{(0.025,14)}\ se_{\hat{\beta}_5}) \end{cases}$$

The results are shown as below:

```r
confint(g)[c(4,6),]
```

```
##             2.5 %       97.5 %
## x3 -3.713929e-05 0.0002927368
## x5 -1.652198e-05 0.0002998305
```

b.

$$CI = ((\hat{\beta}_3 + 2\ \hat{\beta}_5) - t_{(0.025,14)}\ se_{\hat{\beta}_3+2\ \hat{\beta}_5}\ ,\ (\hat{\beta}_3 + 2\ \hat{\beta}_5) + t_{(0.025,14)}\ se_{\hat{\beta}_3+2\ \hat{\beta}_5})$$

$$where\ \ se_{\hat{\beta}_3+2\ \hat{\beta}_5} = \sqrt{\hat{var}(\hat{\beta}_3) + 2^2\ \hat{var}(\hat{\beta}_5) + 4\ \hat{cov}(\hat{\beta}_3\ ,\ \hat{\beta}_5)}\ ,\ and\ \ \hat{cov}(\hat{\beta}_i\ ,\ \hat{\beta}_j) = (X^T X)^{-1}_{ij}\ \hat{\sigma}^2$$

The result is shown as below:

```r
x = model.matrix(g)
xtxi = solve(t(x) %*% x)
sigma = summary(g)$sig
sd_error = sqrt(xtxi[4,4]*sigma^2+4*xtxi[6,6]*sigma^2+4*xtxi[4,6]*sigma^2)
estimate = g$coe[4]+2*g$coe[6]
CI = c(estimate-qt(0.975, g$df)*sd_error, estimate+qt(0.975, g$df)*sd_error)
names(CI) = c("Lower Bound", "Upper Bound")
CI
```

```
##  Lower Bound  Upper Bound
## 5.898666e-05 7.632279e-04
```

c.

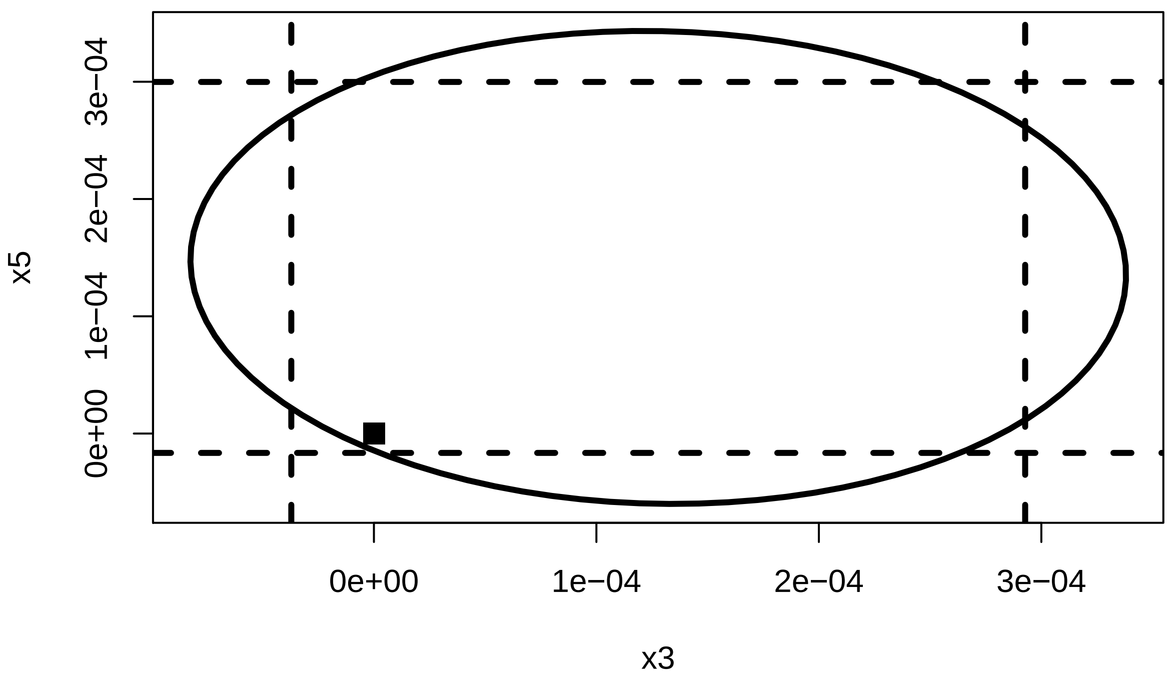
$$\begin{cases} H_0 : \beta_3 \ = \ \beta_5 \ = \ 0 \\ H_1 : at \ \ least \ \ one \ \ of \ \ \beta_3 \ \ or \ \ \beta_5 \ \ \neq \ 0 \end{cases}$$

```
library(ellipse)
```

```
##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs
```

```
plot(ellipse(g, c(4,6)), lwd = 3, type = "l")
points(0,0, cex = 1.5, pch = 15)
abline(v=c(confint(g)[4,1], confint(g)[4,2]),lwd=3,lty=2)
abline(h=c(confint(g)[6,1], confint(g)[6,2]),lwd=3,lty=2)
```



From the above ellipse, the origin is in that ellipse, means $fail \ \ to \ \ reject \ \ H_0$. Therefore, we do not have enough evidence to show that at least one of β_3 or $\beta_5 \ \neq \ 0$ as x_1 , x_2 , x_4 are in the model.

d.

探討 $(\beta_1.\beta_2, \beta_3, \beta_4, \beta_5)$ 所建構的 95% 聯合信賴區間是否包含 $(0, 0, 0, 0, 0)$，同等於在 5% 顯著水準下做以下假設檢定：

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : at \ least \ one \ \beta_i \ \neq \ 0 \ , \ i \ = \ 1, ..., 5 \end{cases}$$

即為進行 full model 的 full test：

```
summary(g)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39447 -0.11847  0.00053  0.08313  0.56232
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.156e+00  9.135e-01  -2.360   0.0333 *
## x1          -9.012e-06  5.184e-04  -0.017   0.9864
## x2           1.316e-03  1.263e-03   1.041   0.3153
## x3           1.278e-04  7.690e-05   1.662   0.1188
## x4           7.899e-03  1.400e-02   0.564   0.5815
## x5           1.417e-04  7.375e-05   1.921   0.0754 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2618 on 14 degrees of freedom
## Multiple R-squared:  0.8107, Adjusted R-squared:  0.743
## F-statistic: 11.99 on 5 and 14 DF,  p-value: 0.0001184
```

$\because p-value \ = \ 0.0001184 \ < \ 0.05 \ \Rightarrow \ reject \ H_0$

$\therefore$ the origin $(0,0,0,0,0)$ would lie inside the 95% confidence region for $(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)$

e.
檢定非揮發性固體 $(x_3 - x_4)$ 對於反映變數是否有線性效應，即判斷模型

$$y = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \beta_3^*\ (x_3 - x_4) + \beta_4^*\ x_4 + \beta_5\ x_5 + \epsilon$$

中係數 β_3^* 是否為 0，而該模型可從原 full model 移項得到

$$y = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \beta_3\ (x_3 - x_4) + (\beta_3 + \beta_4)\ x_4 + \beta_5\ x_5 + \epsilon$$

從此模型可得知，本題所求即為在 full model 下做以下假設檢定

$$\begin{cases} H_0 : \beta_3 \ = \ 0 \\ H_1 : \beta_3 \ \neq \ 0 \end{cases}$$

而 a. 小題中 β_3 的信賴區間有包含 0，所以 fail to reject H_0

$\Rightarrow$ We do not have enough evidence to show that non-volatile solids have linear effect on the response under the full model.

Problem 2.
a.
Construct the full model:

$$lqsa \ = \ \beta_0 + \beta_1\ lcavol + \beta_2\ lweight + \beta_3\ age + \beta_4\ lbph + \beta_5\ svi + \beta_6\ lcp + \beta_7\ gleason + \beta_8\ pgg45 + \epsilon$$

```r
data2 = read.table("prostate.txt", header = T)
g2 = lm(lpsa ~ lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45, data = data2)
```

i.

$$\begin{cases} 90\%\ CI &= (\hat{\beta}_3 - t_{(0.05,88)}\ se_{\hat{\beta}_3}\ ,\ \hat{\beta}_3 + t_{(0.05,88)}\ se_{\hat{\beta}_3}) \\ 95\%\ CI &= (\hat{\beta}_3 - t_{(0.025,88)}\ se_{\hat{\beta}_3}\ ,\ \hat{\beta}_3 + t_{(0.025,88)}\ se_{\hat{\beta}_3}) \end{cases}$$

The results are shown as below:

```r
confint(g2, level = 0.9)[4,]
```

```
##          5 %         95 %
## -0.038210200 -0.001064151
```

```r
confint(g2)[4,]
```

```
##        2.5 %       97.5 %
## -0.041840618  0.002566267
```

```r
summary(g2)$coe[4,]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.01963718  0.01117272 -1.75759949  0.08229321
```

age 的 $90\%\ CI$ 並沒有包含 0，但 $95\%\ CI$ 則有包含 0，可由此推得 $0.05 < p-value < 0.1$，觀察 $regression\ summary$ 報表所呈現的 $p-value = 0.08229321$ 的確符合該條件。

ii.

Let $X_0 = (1.44692, 3.62301, 65, 0.3001, 0, -0.79851, 7, 15)^T$, and $\beta = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7, \hat{\beta}_8)^T$

then the predicted value $Y_0 = X_0^T\ \beta = 2.389053$

the standard error of the predicted value $se_{Y_0} = \sqrt{1 + X_0^T (X^T X)^{-1} X_0}\ \hat{\sigma}$

$\therefore$ The $95\%\ CI = (Y_0 - t_{(0.025,88)}\ se_{Y_0}\ ,\ Y_0 + t_{(0.025,88)}\ se_{Y_0})$

The result is shown as below:

```r
df = data.frame(lcavol = 1.44692,
                lweight = 3.62301,
                age = 65,
                lbph = 0.3001,
                svi = 0,
                lcp = -0.79851,
                gleason = 7,
                pgg45 = 15)
predict(g2, df, interval = "prediction")
```

```
##        fit       lwr      upr
## 1 2.389053 0.9646584 3.813447
```

iii.
Now, the observation has been changed into $X_1 = (1.44692, 3.62301, 20, 0.3001, 0, -0.79851, 7, 15)^T$, and then do the same calculation as above to attain the 95% CI.

The result is been shown as below:

```
df2 = data.frame(lcavol = 1.44692,
                 lweight = 3.62301,
                 age = 20,
                 lbph = 0.3001,
                 svi = 0,
                 lcp = -0.79851,
                 gleason = 7,
                 pgg45 = 15)
predict(g2, df2, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.272726 1.538744 5.006707
```

因為 age 變數數據全部都落在 41~79 的區間之中，所以 $age = 20$ 為外插 (extrapolation) 的資料，在根據此筆數據做估計時，誤差範圍 (即信賴區間寬度) 就會變得較寬。

b.
Construct the reduced model:

$$lqsa \;=\; \beta_0 + \beta_1 \; lcavol + \beta_2 \; lweight + \beta_5 \; svi + \epsilon$$

```
g3 = lm(lpsa ~ lcavol+lweight+svi, data = data2)
```

i.
Use the same method to calculate the prediction and the confidence interval.

```
df3 = data.frame(lcavol = 1.44692,
                  lweight = 3.62301,
                  svi = 0)
predict(g3, df3, interval = "prediction")
```

```
##        fit       lwr      upr
## 1 2.372534 0.9383436 3.806724
```

計算出來的估計值與 part a 差異不大，信賴區間的寬度也只有稍微寬於 part a 所計算出的寬度。

我會更傾向於選擇 part b 的模型，因為使用較少的變數，即代表在時間和金錢上的成本花費較少，而且在此題的情況下，計算出的預測值和信賴區間都沒有太大的差異。

ii.

$$\begin{cases} H_0 : reduced\ model\ fits\ better \\ H_1 : full\ model\ fits\ better \end{cases} \Longleftrightarrow \begin{cases} H_0 : \beta_3 = \beta_4 = \beta_6 = \beta_7 = \beta_8 \\ H_1 : at\ least\ one\ \beta_i \;\neq\; 0\ ,\ i = 3, 4, 6, 7, 8 \end{cases}$$

```
anova(g3, g2)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 47.785
## 2     88 44.163  5    3.6218 1.4434 0.2167
```

$\because p - value \ = \ 0.2167 \ > \ 0.05 \ \Rightarrow \ fail\ to\ reject\ H_0$

$\therefore$ We do not have enough evidence to show that the full model fits better than reduced model. The reduced model is prefered.