1. 14/15

2. 8.5/10

HW 4-Linear Model

ID : 111024517 Name : 鄭家豪

due on 11/17

Problem 1

y : log (oxygen demand, mg oxygen per minute)

x1 : biological oxygen demand, mg/liter

x2 : total Kjeldahl nitrogen, mg/liter

x3 : total solids, mg/liter

x4 : total volatile solids, a component of x3, mg/liter

x5 : chemical oxygen demand, mg/liter

Model : $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i$

,where $i = 1, 2, ..., n$, $\epsilon = (\epsilon_1, ..., \epsilon_n)^T \sim N(\mathbf{0}, \sigma^2 \mathbf{I_n})$.

Parameter space : $\Omega = \{\beta \in \mathbb{R}^6 : y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon\}$

```
data <- read.table('http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/wastes.txt',
                   header=TRUE,fill = TRUE)
colnames(data)[1:7] <- names(data)[2:8]
data = data[-c(21),]
data = data[,-c(8)]
model <- lm(y ~ x1+x2 +x3 +x4 + x5,data = data)
```

2 **(a)**

The 95% confidence interval for β_3 is $\hat{\beta}_3 \pm t_{(n-p,1-0.025)} * s.e.(\hat{\beta}_3)$:

```
coef <- summary(model)$coef
bound_L <- coef[4,1] - qt(0.975,20-6)*coef[4,2]
bound_U <- coef[4,1] + qt(0.975,20-6)*coef[4,2]
kable(t(c(bound_L,bound_U)),col.names = NULL)
```

-3.71e-05	0.0002927

由以上式子，得到 95% C.I. for β_3 is $(-3.713929 \times 10^{-5}, 2.927368 \times 10^{-4})$.
Similarly,the 95% confidence interval for β_5 is $\hat{\beta}_5 \pm t_{(n-p,1-0.025)} * s.e.(\hat{\beta}_5)$:

```
bound_L <- coef[6,1] - qt(0.975,20-6)*coef[6,2]
bound_U <- coef[6,1] + qt(0.975,20-6)*coef[6,2]
kable(t(c(bound_L,bound_U)),col.names = NULL)
```

-1.65e-05	0.0002998

由以上式子，得到 95% C.I. for β_5 is $(-1.652198 \times 10^{-5}, 2.998305 \times 10^{-4})$.
By the way, 這邊可以使用指令:"confint" 來找出每個 β 的信賴區間:

```
confint(model,level = 0.95)
```

```
##                     2.5 %        97.5 %
## (Intercept) -4.115384e+00 -0.1969076588
## x1          -1.120765e-03  0.0011027419
## x2          -1.394016e-03  0.0040257880
## x3          -3.713929e-05  0.0002927368
## x4          -2.212779e-02  0.0379255006
## x5          -1.652198e-05  0.0002998305
```

(b)

The 95% confidence interval for $\beta_3 + 2\beta_5$ is $\hat{\beta}_3 \pm t_{(n-p,1-0.025)} * s.e.(\hat{\beta}_3 + \hat{\beta}_5)$
,where $s.e.(\hat{\beta}_3 + \hat{\beta}_5) = \sqrt{\text{Var}(\hat{\beta}_3) + \text{Var}(2\hat{\beta}_5) + 4\text{Cov}(\beta_3, \beta_5)}$.

$$\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = (X^T X)^{-1}_{ij}\hat{\sigma}^2 \Rightarrow \text{Cov}(\hat{\beta}_3, \hat{\beta}_5) = \hat{\sigma}^2(-2.612349 \times 10^{-9})$$

```
X <- model.matrix(model)
solve(t(X)%*%X)[4,6]
```

```
## [1] -2.612349e-09
```

(以上為 $(X^T X)^{-1}_{35}$ 的數值)
因此，我們可以得到 95% C.I. for $\beta_3 + 2\beta_5$ by the following:

```
sigma_hat <- summary(model)$sig
sd <- sqrt(coef[4,2]^2+4*coef[6,2]^2 + 4*(-2.612349e-09)*sigma_hat^2)
bound_L <- coef[4,1] + 2*coef[6,1] - qt(0.975,20-6)*sd
bound_U <- coef[4,1] + 2*coef[6,1] + qt(0.975,20-6)*sd
kable(t(c(bound_L,bound_U)),col.names = NULL)
```

5.9e-05	0.0007632

95% C.I. for $\beta_3 + 2\beta_5$ is $(5.898666 \times 10^{-5}, 7.632279 \times 10^{-4})$.

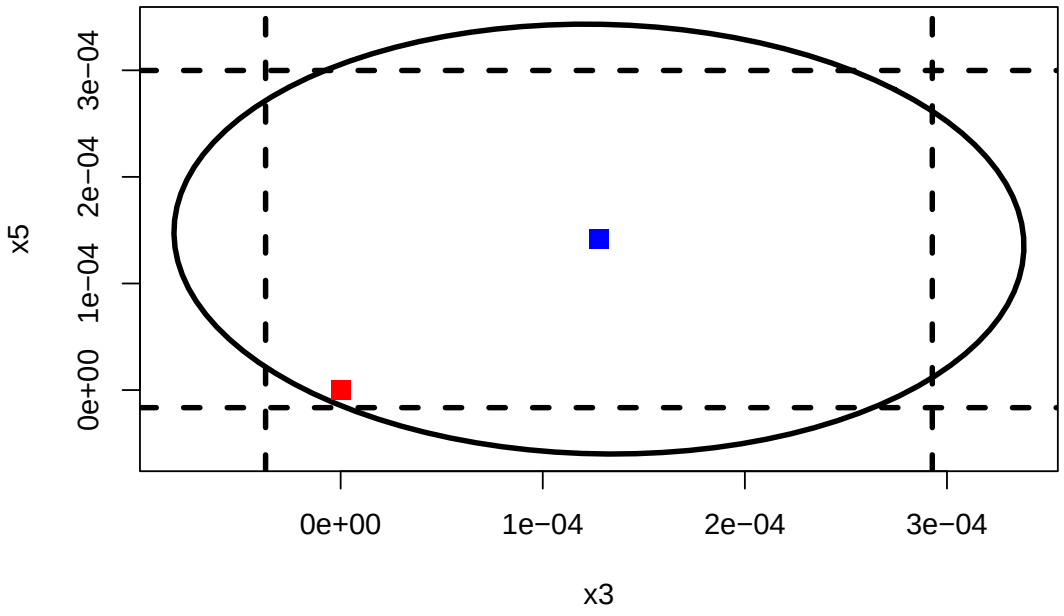
(c)

紅點:(0,0) ; 藍點:$(\hat{\beta}_3, \hat{\beta}_5)$

虛線 (x-axis):95% 信賴區間 for β_3

虛線 (y-axis):95% 信賴區間 for β_5

```
library(ellipse)
plot(ellipse(model,c(4,6)),lwd=3, type="l")
points(model$coef[4],model$coef[6],
       cex = 1.5,pch=15,col="blue")
points(0,0,cex=1.5,pch=15,col= "red")
abline(v=confint(model,level = 0.95)[4,],
       lwd=3,lty=2)
abline(h=confint(model,level = 0.95)[6,],
       lwd=3,lty=2)
```



由 (a) (b) 可以得知，β_3 與 β_5 的信賴區間分別都會涵蓋 0 ，這代表著分別做檢定:

$$\mathrm{H}_0 : \beta_i = 0 \quad \text{v.s.} \quad \mathrm{H}_1 : \beta_i \neq 0$$

都會得到這個結論: “Do not reject H_0 at significant level 0.05”。那如果要檢定:

$$H_0 : \omega = \{\beta \in \mathbb{R}^6 : \beta_3 = \beta_5 = 0\} \text{ v.s. } H_1 : \Omega/\omega$$

可觀察上面的 confidence region for (β_3, β_5)，是否有涵蓋到原點 $(0, 0)$。這裡可以觀察到，此圖是會涵蓋原點！因此，Do not reject H_0 at significant level 0.05。

2

(d)

The confidence region of $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ is satisfying the following:

$$(A(\hat{\beta} - \beta))^T (A(X^T X)^{-1} A^T)^{-1} (A(\hat{\beta} - \beta)) \leq (5\hat{\sigma}^2) F_{5,20-6}(\alpha) \text{ under } H_0 - (*)$$

which implies the testing :

$$H_0 : \omega = \{\beta : (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = \mathbf{0}\} \text{ v.s. } H_1 : \Omega/\omega$$

,where $A = [\mathbf{0}_{5\times 1} | \mathbf{I}_{5\times 5}]$:$5 \times 6$ matrix, is equivalent to under the null hypothesis is true, examine whether $(*)$ holds.
So,by the following,

```
A = matrix(0,nrow = 5,ncol = 6)
for (i in 1:5){
  A[i,i+1] =1
}
model_inverse <- solve(t(X)%*%X)
statistic <- as.numeric(t(A%*%coef[,1]) %*% solve(A%*%model_inverse%*%t(A)) %*% A%*%(coef[,1]))
critical_value <- 5*summary(model)$sig^2*qf(0.95,5,20-6)
statistic - critical_value
```

```
## [1] 3.094621
```

計算 $(\hat{A\beta} - 0)^T (A(X^T X)^{-1} A^T)^{-1} (\hat{A\beta} - 0) - (5\hat{\sigma}^2) F_{5,20-6}(\alpha = 0.05) = 3.094621 > 0$，代表 $(*)$ 不成立，即得到結論:“Reject H_0 at significant level 0.05”。

Does origin lie inside or outside of confidence region?

5

(e)

let total non-volatile solids be $X_6 = X_3 - X_4$
Model : $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{4i} + X_{6i}) + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i$ 在進行檢定前，我們發現 $\text{Cor}(X_3, X_6) = 0.9999964 \approx 1$，這代表 X_3 和 X_6 存在共線性的性質。

8.5

```
cor(data$x3,data$x3-data$x4)
```

```
## [1] 0.9999964
```

這意味著 X_6 的效應和 X_3 極度相似，so, the testing to this suspicion :

$$H_0 : \omega = \{\beta : \beta_3 = 0\} \quad \text{v.s.} \quad H_1 : \Omega/\omega$$

這裡的檢定等價於利用 β_3 的信賴區間來檢驗是否涵蓋 0。由 (a) 可以得知，其區間是涵蓋 0 的，因此，Do not reject H_0 at significant level 0.05。

Problem 2

0.5

(a)

```
data <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/houseprices.txt",
                   header=TRUE)
summary(data)
```

```
##      Price            SQFT           Age           Features          NE
##  Min.   : 540   Min.   : 837   Min.   : 1.00   Min.   :0.00   Min.   :0.0000
##  1st Qu.: 780   1st Qu.:1280   1st Qu.: 5.75   1st Qu.:3.00   1st Qu.:0.0000
##  Median : 960   Median :1549   Median :13.00   Median :4.00   Median :1.0000
##  Mean   :1063   Mean   :1654   Mean   :14.97   Mean   :3.53   Mean   :0.6667
##  3rd Qu.:1200   3rd Qu.:1894   3rd Qu.:19.25   3rd Qu.:4.00   3rd Qu.:1.0000
##  Max.   :2150   Max.   :3750   Max.   :53.00   Max.   :8.00   Max.   :1.0000
##                                NA's   :49
##      Corner           Tax
##  Min.   :0.000   Min.   : 223.0
##  1st Qu.:0.000   1st Qu.: 600.0
##  Median :0.000   Median : 731.0
##  Mean   :0.188   Mean   : 793.5
##  3rd Qu.:0.000   3rd Qu.: 919.0
##  Max.   :1.000   Max.   :1765.0
##                  NA's   :10
```

由以上的 summary table，可以發現 Age 的 NA 數量大約佔了總樣本數的四成 $(49/117 \approx 0.419)$，可能因為這組數據的房子，有很多陳舊已久的房子以至於其房屋年齡無從而知，加上非 NA 的 Age 樣本中，大約有 75% 的房屋年齡在 20 年以下，代表大部分採樣的年齡屬於年輕的一群，假如 NA 代表陳舊已久的房屋歲數的話，那只移除有 NA 的 row data 來做分析，主要只會分析年輕房子的售價！因此，Age 這個 predictor 無具代表性，應擇將其 predictor 移除會比較保險。

3.

(b)

由 (a) 的 summary teble, 會發現 Taxes 有 10 個 NA 值，由於 NA 佔的比例不高 (10/117<0.1)，另外考量 Tax 為 NA 的可能性為房主逃稅，這種類型的房子不太值得去做分析，因此這裡將其 10 筆 row data 給移除掉。於是整理以上，資料整合為:

```
data = data[,-c(3)]
data = data[-c(which(is.na(data$Tax))),]
```

整合完後的資料為 non-NA data with size: 107 x 6

```
model <- lm(Price ~ . ,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -544.22  -74.05  -15.03   68.34  615.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  77.6954    62.1231   1.251   0.2139
## SQFT          0.2666     0.0620   4.301 3.93e-05 ***
## Features     13.8581    13.5727   1.021   0.3097
## NE           -3.3995    36.4875  -0.093   0.9260
## Corner      -89.1245    42.4246  -2.101   0.0382 *
## Tax           0.6627     0.1097   6.042 2.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 171.9 on 101 degrees of freedom
## Multiple R-squared:  0.8091, Adjusted R-squared:  0.7996
## F-statistic:  85.6 on 5 and 101 DF,  p-value: < 2.2e-16
```

這裡得到模型:

$\hat{y}_{\text{Price}} = 77.6954 + 0.2666X_{\text{SQFT}} + 13.8581X_{\text{Features}} - 3.3995X_{\text{NE}} - 89.1245X_{\text{Corner}} + 0.6627X_{\text{Tax}}$

The 100(1-α)% confidence interval for β_i is $\hat{\beta}_i \pm t_{(107-6,1-\alpha)} * \text{s.e.}(\hat{\beta}_i)$,by 指令"confint"，可以分別得到 95%,99% 的信賴區間:

```
confint(model,level=0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept)  -45.5400249 200.9309248
## SQFT           0.1436693   0.3896326
## Features     -13.0665047  40.7827060
## NE           -75.7809282  68.9818807
## Corner      -173.2835186  -4.9654323
## Tax            0.4451310   0.8803222
```

```
confint(model,level=0.99)
```

```
##                    0.5 %      99.5 %
## (Intercept)  -85.4016058 240.7925057
## SQFT           0.1038899   0.4294120
## Features     -21.7755013  49.4917026
## NE           -99.1933199  92.3942724
## Corner      -200.5054895  22.2565386
## Tax            0.3747479   0.9507053
```

從 summary of lm，可以看到 Corner 的 p-value =0.0382 $>$ 0.01 但是 $<$ 0.05。對於此檢定:

$$\mathrm{H}_0 : \beta_{\text{Corner}} = 0 \quad \text{v.s.} \quad \mathrm{H}_1 : \beta_{\text{Corner}} \neq 0$$

如果 p-value 小於給定的顯著水準 α，應拒絕 H_0，反之則接受 H_0。故使用 p-value 來檢定 $\beta_{\text{Corner}} = 0$ 與檢察其係數之信賴區間是否涵蓋 0，兩者是等價的。從上面的信賴區間，95% C.I.:(-173.2835186,-4.9654323) 沒有涵蓋 0，但 99% C.I.:(-200.5054895,22.2565386) 有涵蓋 0，與 p-value 推出的檢定結果一致。

(c)

這題非預測這類型房屋的平均價格，應為對 future observation 的房價預測，因此其預測的信賴區間 (預測區間) for Y_{new} at $x_0 = (1, SQFT = 2500, Features = 5, NE = 1, Corner = 1, Tax = 1200)^T$ 為:

$$\hat{Y}_{\text{new}} \pm t_{(20-6,1-\alpha/2)} \times \hat{\sigma}\sqrt{1 + x_0^T (X^T X)^{-1} x_0}, \text{ where } X : \text{model matrix}$$

可使用指令 predict 來計算其區間:

```
predict(model,data.frame(SQFT = 2500,Features=5,NE=1,
                         Corner = 1,Tax=1200),
        se=TRUE,interval = "prediction")
```

```
## $fit
##        fit      lwr      upr
## 1 1516.361 1162.761 1869.961
##
## $se.fit
## [1] 47.20802
##
## $df
## [1] 101
##
## $residual.scale
## [1] 171.8849
```

2

這裡預估這間房子的售價為 1516.361 (百美元)，以及其 95% 預測區間為 (1162.761,1869.961)，即有 95% 的信心預估其價格介於 (1162.761,1869.961) 之間。

(d)

由於這裡已知的資訊只有 SQFT = 2500，這裡我採用兩個方式來做預測並分別做分析:

- 內插 (imterpolation)：對 Features、NE、Corner 與 Tax 進行內插，使用 (b) 的模型來做預測需要內插的值應落在資料涵蓋的範圍內，這樣預測結果才不太會受配適模型中不顯著效應變數的影響。由 (a) 的 summary table 可以觀察到，Tax 存在離群值且有 6 個。

3

```
Q1 <- as.numeric(quantile(data$Tax,0.25))
Q3 <- as.numeric(quantile(data$Tax,0.75))
length(which(data$Tax>Q3+1.5*(Q3-Q1)))
```

```
## [1] 6
```

為了避免受到離群值的影響，Tax 的內插值設定為其中位數 731。另外，其餘 predictors 為類別型資料，內插值分別皆設定為其眾數。

```
Mode <- function(x) {
  ux <- unique(x)
  tab <- tabulate(match(x, ux))
  ux[tab == max(tab)]
}
predict(model,data.frame(SQFT = 2500,
                         Features=Mode(data$Features),
                         NE=Mode(data$NE),
                         Corner = Mode(data$Corner),
```

```
                          Tax=median(data$Tax)),
        se=TRUE,interval = "prediction")
```

```
## $fit
##        fit      lwr      upr
## 1 1280.809 917.6289 1643.989
##
## $se.fit
## [1] 63.0362
##
## $df
## [1] 101
##
## $residual.scale
## [1] 171.8849
```

最後我們得到預估值為 1280.809(百美元)。這裡要注意的是，因為內插是估計未知 predictors 的值，代表會產生隨機效應，所以計算出來的 95% 預測區間，其 coverage probability 實際上會大於 0.95，這不符合我們題目所要找的 95% 預測區間。

- Predictor 只保留 SQFT，然後配適模型: $y_{\text{Price}} = \beta_0 + \beta_1 X_{\text{SQFT}} + \epsilon$

```
model <- lm(Price~ SQFT,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ SQFT, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1050.93   -93.54     1.44    60.58   749.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.82322   66.39183   0.931    0.354
## SQFT         0.60910    0.03796  16.047   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 207.6 on 105 degrees of freedom
```

```
## Multiple R-squared:  0.7103, Adjusted R-squared:  0.7076
## F-statistic: 257.5 on 1 and 105 DF,  p-value: < 2.2e-16
```

由以上，我們得到配適模型: $\hat{y}_{\text{Price}} = 61.82322 + 0.60910 \times X_{\text{SQFT}}$。
利用此模型來做預測:

```
predict(model,data.frame(SQFT = 2500),
                         se=TRUE,
                         interval = "prediction")
```

```
## $fit
##        fit      lwr      upr
## 1 1584.563 1166.205 2002.921
##
## $se.fit
## [1] 37.4438
##
## $df
## [1] 105
##
## $residual.scale
## [1] 207.6429
```

在只知道 SQFT=2500 的資訊下，預估這間房子的售價為 1584.563(百美元)，由於這裡沒有考慮其他 predictors 的隨機性，所以其 95% 預測區間為 (1166.205,2002.921)，即有 95% 的信心預估其價格介於 (1166.205,2002.921) 之間。
因此，總結上述，預測值 = 1584.563，其 95% 預測的信賴區間為 (1166.205,2002.921)。