1. 10/10

2. 8/10

3. 10/10

# HW2-Linear model

ID : 111024517　　　　Name : 鄭家豪

due on 10/20 (Tue)

**1.**

先讀取資料來檢查資料性質 (這裡我採用讀取前 10 筆資料):

| Output <int> | SI <int> | SP <int> | I <dbl> |
|---|---|---|---|
| 12090 | 56 | 840 | 10.54 |
| 11360 | 133 | 2040 | 11.11 |
| 12930 | 256 | 2410 | 10.73 |
| 12590 | 382 | 2760 | 14.29 |
| 16680 | 408 | 2520 | 11.19 |
| 23090 | 572 | 2950 | 14.03 |
| 16390 | 646 | 2480 | 18.76 |
| 16180 | 772 | 2270 | 13.53 |
| 17940 | 805 | 4040 | 16.71 |
| 18800 | 919 | 2750 | 14.74 |

根據每個變數的值變動情況，每個變數皆可視為量化型變數。

## 3 (a)

- Response variab: Output($Y$)
- Explantary variables: SI($X_1$)、SP($X_2$)、I($X_3$)

  Assumptions:

  $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i$,where $E(\epsilon_i) = 0$ , $Var(\epsilon_i) = \sigma^2$ for i=1,2,...,17.

  $Model\ matrix:\ X = [1, X_1, X_2, X_3]$

  $Coefficient: \beta =\ [\beta_0, \beta_1, \beta_2, \beta_3]$

The Least square method : $\hat{\beta} = (X^T X)^{-1} X^T Y$

```
X <- data.matrix(cbind(rep(1,17),ex1_data[,2:4]))
Y <- ex1_data[,1]
round(solve(t(X) %*% X) %*% t(X) %*% Y,digits = 4)
```

計算以上算式，會得到 $[6026.0607, 1.7422, 5.3019, -255.5056]^T$ (round to 4 decimal places)

因此藉由 Least squares，得到 $\hat{Y} = 6026.0607 + 1.7422X_1 + 5.3019X_2 - 255.5056X_3$

## 3 (b)

Denote $SI^2$ and $SP \times I$ be $X_4$ and $X_5$,respectively.

Assumptions:

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i$

,where $E(\epsilon_i) = 0$ , $Var(\epsilon_i) = \sigma^2$ for i=1,2,...,17.

$Model\ matrix:\ X^* = [1, X_1, X_2, X_3, X_4, X_5]$

$Coefficient: \beta^* =\ [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5]$

一樣藉由 Least square method，計算 $\hat{\beta}^* = ((X^*)^T X^*)^{-1} (X^*)^T Y$，

```
X_star <- data.matrix(cbind(X,ex1_data[,2]^2,ex1_data[,3]*ex1_data[,4]))
round(solve(t(X_star) %*% X_star) %*% t(X_star) %*% Y,digits = 4)
```

會得出 $[52404.5295, 35.1319, -13.7152, -3715.9028, -0.0145, 1.0221]^T$ (round to 4 decimal places)

因此得到 $\hat{Y} = 52404.5295 + 35.1319X_1 - 13.7152X_2 - 3715.9028X_3 - 0.0145X_4 + 1.0221X_5$.

4 **(c)**

由於 $X_4 = X_1^2, X_5 = X_2 * X_3$，

可將 Part(b) 的結果另為 $F(X_1, X_2, X_3)$ : the function of $(X_1, X_2, X_3)$。

因為解釋變數之間不相互影響，不過 $X_5$ 是 $X_2$ 和 $X_3$ 的交互作用變數，所以可再將其拆成

$F(X_1, X_2, X_3) = F_1(X_1) + F_2(X_2, X_3)$

,where $F_1(X_1) = 52404.5295 + 35.1319X_1 - 0.0145X_1^2$ ; $F_2(X_2, X_3) = -13.7152X_2 - 3715.9028X_3 + 1.0221X_2X_3$.

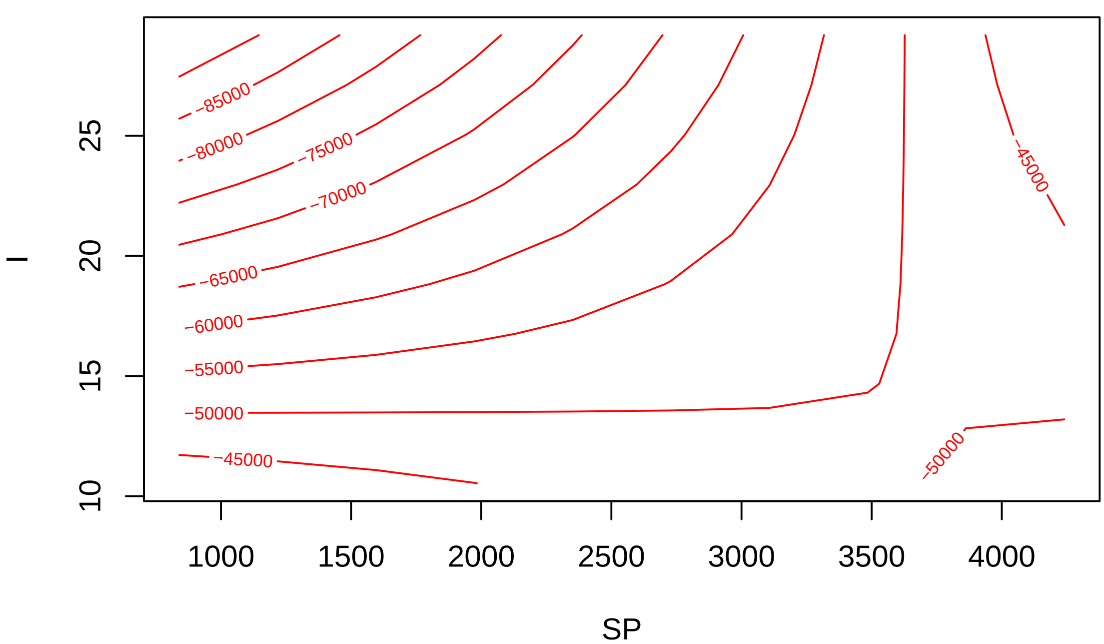
(i) 由於 $F_1$ 是凹函數，所以由”Second derivative test” 得知:

$$\frac{d}{dx}F_1(x) = 0$$
$$\Rightarrow 35.1319 - 0.029x = 0 \Rightarrow x = 1211.445$$

考慮 $SI$ 的資料範圍為 [56,1754]，當 $X_1 = 1211$ 時 (因為 $SI$ 皆為整數)，$F_1(X_1)$ 達至最大值 73684.71，這是合理的。

(ii) $F_2$ 是雙變數函數，由於 determinant of Hessian matrix <0 for all $X_2, X_3$ ，在找最大值時可考慮其邊界點。

所以這裡先觀察 $F_2$ 的 contour plot:




這裡可以觀察到，(i) SP 極大且 I 相對極大或 (ii) SP 極小且 I 相對極小時，$F_2$ 可以達至最大。

For case (i)，我使用 R 語言的"optim" 指令來解最大值 (初始值分別設定為 $SP = P_{90} = 3938$(90-th percentile) 和 $I = P_{90} = 25.944$)。

```
F_2=function(x){-(-13.7152*x[1]-3715.9028*x[2]+1.0221*x[1]*x[2])}
optim(c(3938,25.944),fn = F_2,method = 'L-BFGS-B',lower = c(840,10.54),upper = c(4240,29.19))
```

```
$par
[1] 4240.00   29.19

$value
[1] 40118.83
```

**NOTE: R** 語言的 **optim** 指令為求解函數最小值，因此使用指令時要在函數前多個負號。

For case (ii)，初始值分別設定為 $SP = P_{10} = 2178$ 和 $I = P_{10} = 10.958$:

```
optim(c(2178,10.958),fn = F_2,method = 'L-BFGS-B',lower = c(840,10.54),upper = c(4240,29.19))
```

```
$par
[1] 840.00  10.54

$value
[1] 41637.12
```

由上述結果，我們得知 $SP = 4240$ , $I = 29.19$ 可以得到 $F_2$ 的最大值 $F_2(4240, 29.19) = -40118.83$。

因此，the maximum of $\hat{Y} = 52404.5295 + 35.1319X_1 - 13.7152X_2 - 3715.9028X_3 - 0.0145X_4 + 1.0221X_5$ is $73684.71 - 40118.83 = 33565.88$ at $(SI, SP, I) = (1211, 4240, 29.19)$。

**2.**

- 讀取資料

```
ex2_data <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/prostate.txt"
,header = TRUE)
```

Denote these variables as the following:
* Response variabe: lpsa($Y$)
* Explantary variables: lcavol($X_1$)、lweight($X_2$)、age($X_3$)、lbph($X_4$)、svi($X_5$)、lcp($X_6$)、gleason($X_7$)、pgg45($X_8$)

### (a)

1. model?

```
a_fit <- lm(lpsa ~ lcavol,data = ex2_data)
```

| beta0 | beta1 | R squard | Residual s.d. |
|---|---|---|---|
| 1.507298 | 0.7193201 | 0.5394319 | 0.7874994 |

### (b)

3

Model adding lweight($X_2$): $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$,i=1,2,...,97.

```
b1_fit <- lm(lpsa ~ lcavol + lweight,data = ex2_data)
```

| beta0 | beta1 | beta2 | R squard | Residual s.d. |
|---|---|---|---|---|
| -0.3026179 | 0.6775253 | 0.5109495 | 0.5859345 | 0.7506469 |

Model adding lweight($X_2$) and svi($X_5$): $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 X_5 + \epsilon_i$,i=1,2,...,97.

```
b2_fit <- lm(lpsa ~ lcavol + lweight + svi,data = ex2_data)
```

| beta0 | beta1 | beta2 | beta5 | R squard | Residual s.d. |
|---|---|---|---|---|---|
| -0.2680926 | 0.551638 | 0.5085413 | 0.6661584 | 0.6264403 | 0.7168094 |

Model adding lweight ,svi, lbph($X_4$) and svi($X_5$):$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i$,i=1,2,...,97.

```
b3_fit <- lm(lpsa ~ lcavol + lweight + lbph + svi,data = ex2_data)
```

| beta0 | beta1 | beta2 | beta4 | beta5 | R squard | Residual s.d. |
|---|---|---|---|---|---|---|
| 0.1455407 | 0.5496031 | 0.3908759 | 0.0900933 | 0.711737 | 0.6366035 | 0.7108232 |

Model adding lweight ,svi ,lbph and age($X_3$):
$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i$,i=1,2,...,97.

```
b4_fit <- lm(lpsa ~ lcavol + lweight + age + lbph + svi,data = ex2_data)
```

| beta0 | beta1 | beta2 | beta3 | beta4 | beta5 | R squard | Residual s.d. |
|---|---|---|---|---|---|---|---|
| 0.9509974 | 0.565608 | 0.423692 | -0.0148923 | 0.1118399 | 0.720955 | 0.6441024 | 0.7073054 |

Model adding lweight ,svi ,lbph, age and lcp($X_6$):
$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon_i$,i=1,2,...,97.

```
b5_fit <- lm(lpsa ~ lcavol + lweight + age + lbph + svi +lcp,data = ex2_data)
```

| beta0 | beta1 | beta2 | beta3 | beta4 |
|---|---|---|---|---|
| 0.9348684 | 0.5876467 | 0.4180838 | -0.0151124 | 0.1138122 |

| beta5 | beta6 | R squard | Residual s.d. |
|---|---|---|---|
| 0.7825645 | -0.0411838 | 0.645113 | 0.7102135 |

Model adding lweight ,svi ,lbph, age,lcp and pgg45($X_8$):

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 + \epsilon_i$,i=1,2,...,97.

```
b6_fit <- lm(lpsa ~ lcavol + lweight + age + lbph + svi +lcp+pgg45,data = ex2_data)
```

| beta0 | beta1 | beta2 | beta3 | beta4 |
|---|---|---|---|---|
| 0.953926 | 0.5916145 | 0.4482924 | -0.0193365 | 0.1076711 |

| beta5 | beta6 | beta8 | R squard | Residual s.d. |
|---|---|---|---|---|
| 0.7577335 | -0.1044823 | 0.0053177 | 0.6544317 | 0.7047533 |

Model adding all explantary variables:

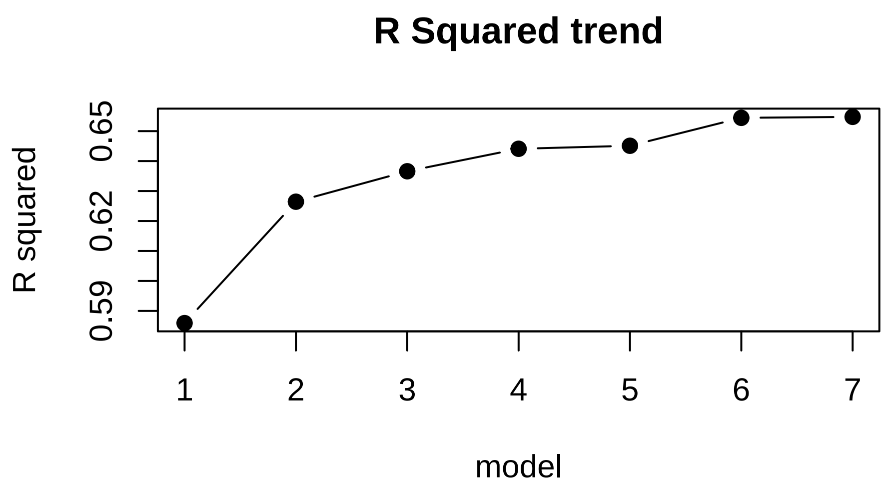
$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \epsilon_i$,i=1,2,...,97.

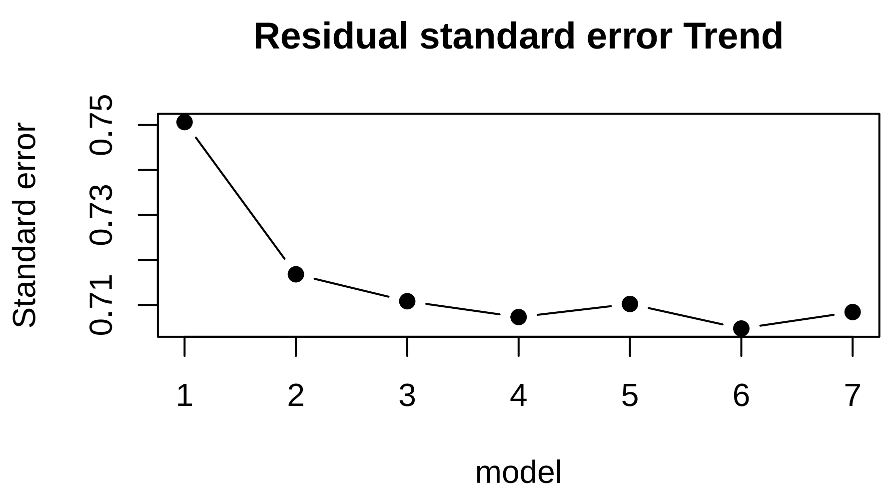
```
b7_fit <- lm(lpsa ~ . ,data = ex2_data)
```

| beta0 | beta1 | beta2 | beta3 | beta4 | beta5 |
|---|---|---|---|---|---|
| 0.6693367 | 0.5870218 | 0.4544674 | -0.0196372 | 0.107054 | 0.7661573 |

| beta6 | beta7 | beta8 | R squard | Residual s.d. |
|---|---|---|---|---|
| -0.1054743 | 0.0451416 | 0.0045252 | 0.6547541 | 0.7084155 |

接著把前面計算的 $R^2$ 與 residual s.d. 值繪製成折線圖



由以上的趨勢圖，

可以發現 part(a) 的模型加進 lweight 和 svi 後的 $R^2$ 值，比單純加 lweight 有明顯的提升，且 residaual standard error 有明顯的降低。

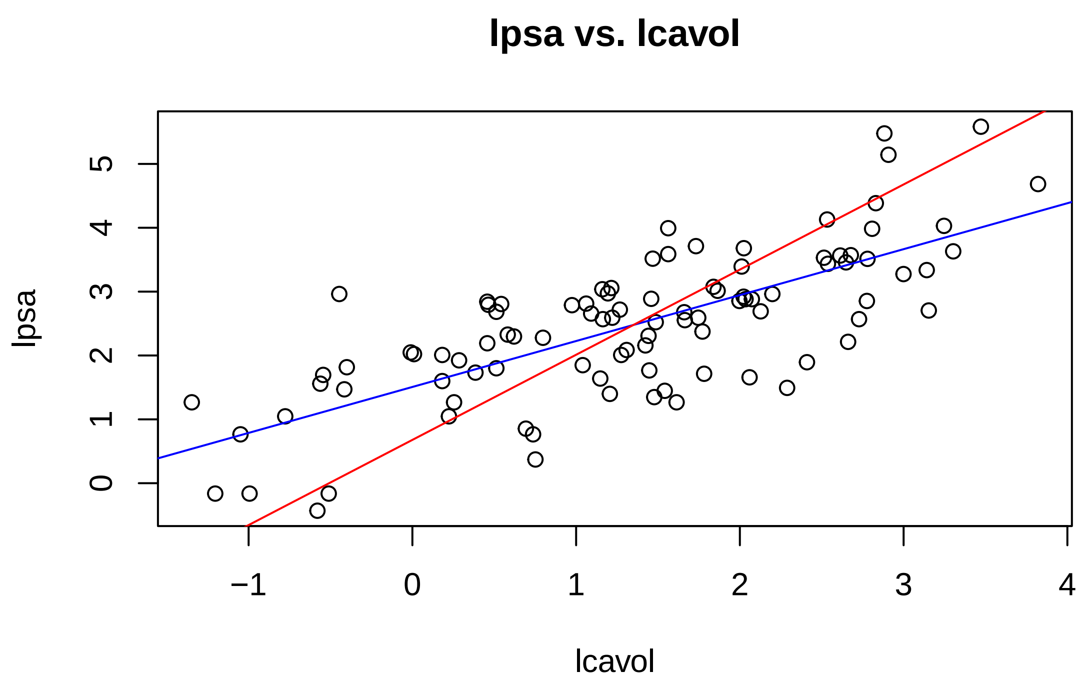
後續加入其他變數，雖然一定不會降低前一個 model 的 $R^2$ 值，但增加的幅度沒有比從 1 -> 2 還要多。

以及 residaual standard error 在 2 之後變化幅度不大。 ?

推測出 lweight 和 svi 同時考慮時，與 lpsa 的變化與解釋性是較顯著相關的。

### (c)

4




藍線: the fitted line of simple regression of lpsa on lcavol,$\hat{Y} = 1.5072979 + 0.7193201X_1$

紅線: the fitted line of simple regression of lcavol on lpsa,$\hat{X}_1 = -0.5085802 + 0.7499191Y$

這裡要注意的是，因為 fitted line of lcavol on lpsa 是以 lcavol 作為反應變數，所以紅線需要做線性轉換才能呈現出來。亦即:

$\hat{X}_1 = \hat{\beta}_0 + \hat{\beta}_1 Y$

$\Rightarrow Y = -\dfrac{\hat{\beta}_0}{\hat{\beta}_1} + \dfrac{1}{\hat{\beta}_1}\hat{X}_1$

$\Rightarrow Y = 0.67818 + 1.333477\hat{X}_1$

這裡可以觀察到，兩條線有一交點。

因為 $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1\overline{X}$ 的性質，不失一般性，紅線也會通過 $(\overline{X}, \overline{Y})$，因此交點為 $(\overline{X}_1, \overline{Y}) = (1.35001, 2.478387)$。

**3.**

- 事前工作: 將資料轉成可讀取的形式

| YEAR | Capital20 | Capital36 | Capital37 | Labor20 | Labor36 | Labor37 | RealValueAdded20 | RealValueAdded36 | RealValueAdded37 |
|---|---|---|---|---|---|---|---|---|---|
| 72 | 243462 | 291610 | 1209188 | 708014 | 881231 | 1259142 | 6496.96 | 6713.75 | 11150.0 |
| 73 | 252402 | 314728 | 1330372 | 699470 | 960917 | 1371795 | 5587.34 | 7551.68 | 12853.6 |
| 74 | 246243 | 278746 | 1157371 | 697628 | 899144 | 1263084 | 5521.32 | 6776.40 | 10450.8 |
| 75 | 263639 | 264050 | 1070860 | 674830 | 739485 | 1118226 | 5890.64 | 5554.89 | 9318.3 |
| 76 | 276938 | 286152 | 1233475 | 685836 | 791485 | 1274345 | 6548.57 | 6589.67 | 12097.7 |
| 77 | 290910 | 286584 | 1355769 | 678440 | 832818 | 1369877 | 6744.80 | 7232.56 | 12844.8 |
| 78 | 295616 | 280025 | 1351667 | 667951 | 851178 | 1451595 | 6694.19 | 7417.01 | 13309.9 |
| 79 | 301929 | 279806 | 1326248 | 675147 | 848950 | 1328683 | 6541.68 | 7425.69 | 13402.3 |
| 80 | 307346 | 258823 | 1089545 | 658027 | 779393 | 1077207 | 6587.33 | 6410.91 | 8571.0 |
| 81 | 302224 | 264913 | 1111942 | 627551 | 757462 | 1056231 | 6746.77 | 6263.26 | 8739.7 |
| 82 | 288805 | 247491 | 988165 | 609204 | 664834 | 947502 | 7278.30 | 5718.46 | 8140.0 |
| 83 | 291094 | 246028 | 1069651 | 604601 | 664249 | 1057159 | 7514.78 | 5936.93 | 10958.4 |
| 84 | 285601 | 256971 | 1191677 | 601688 | 717273 | 1169442 | 7539.93 | 6659.30 | 10838.9 |
| 85 | 292026 | 248237 | 1246536 | 584288 | 678155 | 1195255 | 8332.65 | 6632.67 | 10030.5 |
| 86 | 294777 | 261943 | 1281262 | 571454 | 670927 | 1171664 | 8506.37 | 6651.02 | 10836.5 |

- 讀取資料:

```r
ex3_data <- read.table("ex3.txt",header = TRUE)
```

接著把不同部門的相同變數合併成同一欄 (column)

```r
YEAR <- rep(ex3_data[,1],3)
Capital <- c(ex3_data[,2],ex3_data[,3],ex3_data[,4])
Labor <- c(ex3_data[,5],ex3_data[,6],ex3_data[,7])
RealValueAdded <-c(ex3_data[,8],ex3_data[,9],ex3_data[,10])
Sector <- c(rep("Food and kindred products (20)",15),
            rep("electrical and electronic machinery, equipment and supplies (36)",15),
            rep("transportation equipment (37)",15))
ex3_data <- data.frame(
  "YEAR" = YEAR,"Capital"=Capital,"Labor"=Labor,
  "RealValueAdded"=RealValueAdded,
  "Sector" = Sector
)
```

調整完的資料會變成以下 (取前 6 筆):

| YEAR | Capital | Labor | RealValueAdded | Sector |
|---|---|---|---|---|
| 72 | 243462 | 708014 | 6496.96 | Food and kindred products (20) |
| 73 | 252402 | 699470 | 5587.34 | Food and kindred products (20) |
| 74 | 246243 | 697628 | 5521.32 | Food and kindred products (20) |
| 75 | 263639 | 674830 | 5890.64 | Food and kindred products (20) |
| 76 | 276938 | 685836 | 6548.57 | Food and kindred products (20) |
| 77 | 290910 | 678440 | 6744.80 | Food and kindred products (20) |

## (a)

$V_t = \alpha K_t^{\beta_1} L_t^{\beta_2} \epsilon_t$

$\Rightarrow log(V_t) = log(\alpha K_t^{\beta_1} L_t^{\beta_2} \epsilon_t) = log(\alpha) + \beta_1 log(K_t) + \beta_2 log(L_t) + log(\epsilon_t)$

這裡發現，將其模型取 log 後，就變成是一個線性模型，其變數為:

* Response : $log(V_t)$
* Explantary : $log(K_t)$、$log(L_t)$
* constant : $log(\alpha)$
* error : $log(\epsilon_t)$

```
fit20 <- lm(log(RealValueAdded)~log(Capital)+log(Labor),
            data = ex3_data,
            subset = (Sector =="Food and kindred products (20)"))
beta_20 <- fit20$coefficients[2:3]

fit36 <- lm(log(RealValueAdded)~log(Capital)+log(Labor),
            data = ex3_data,
            subset = (Sector =="electrical and electronic machinery, equipment and supplies (36)"))
beta_36 <- fit36$coefficients[2:3]

fit37 <- lm(log(RealValueAdded)~log(Capital)+log(Labor),
            data = ex3_data,
            subset = (Sector =="transportation equipment (37)"))
beta_37 <- fit37$coefficients[2:3]
result_beta <- matrix(c(beta_20,beta_36,beta_37),ncol = 3)
colnames(result_beta) <- c("Sector.20","Sector.36","Sector.37")
rownames(result_beta) <- c("beta1","beta2")
knitr::kable(result_beta)
```

| | Sector.20 | Sector.36 | Sector.37 |
|---|---|---|---|
| beta1 | 0.2268538 | 0.5260689 | 0.5056509 |
| beta2 | -1.4584782 | 0.2543206 | 0.8454644 |

3.

### (b)

Constraint : $\beta_1 + \beta_2 = 1$

$log(V_t) = log(\alpha) + \beta_1 log(K_t) + (1-\beta_1)log(L_t) + log(\epsilon_t)$

$= log(\alpha) + \beta_1(log(\frac{K_t}{L_t})) + log(L_t) + log(\epsilon_t)$

將 $log(\frac{K_t}{L_t})$ 當作解釋變數，以及 offset 為 $log(L_t)$。

```r
fit20 <- lm(log(RealValueAdded)~log(Capital/Labor),
            data = ex3_data,
            subset = (Sector =="Food and kindred products (20)"),
            offset = log(Labor))
beta_20 <- c(fit20$coefficients[2],1-fit20$coefficients[2])

fit36 <- lm(log(RealValueAdded)~log(Capital/Labor),
            data = ex3_data,
            subset = (Sector =="electrical and electronic machinery, equipment and supplies (36)"),
            offset = log(Labor))
beta_36 <- c(fit36$coefficients[2],1-fit36$coefficients[2])

fit37 <- lm(log(RealValueAdded)~log(Capital/Labor),
            data = ex3_data,
            subset = (Sector =="transportation equipment (37)"),
            offset = log(Labor))
beta_37 <- c(fit37$coefficients[2],1-fit37$coefficients[2])
result_beta <- matrix(c(beta_20,beta_36,beta_37),ncol = 3)
colnames(result_beta) <- c("Sector.20","Sector.36","Sector.37")
rownames(result_beta) <- c("beta1","beta2")
knitr::kable(result_beta)
```

| | Sector.20 | Sector.36 | Sector.37 |
|---|---|---|---|
| beta1 | 1.2896953 | 0.9000888 | 0.0096089 |
| beta2 | -0.2896953 | 0.0999112 | 0.9903911 |

3.

### (c)

$log(V_t) = log(\alpha) + log(\gamma)t + \beta_1 log(K_t) + \beta_2 log(L_t) + log(\epsilon_t)$

將 YEAR 當作解釋變數考慮進來，意味著 Real value added 會隨著年份有所變動。

```
fit20 <- lm(log(RealValueAdded)~YEAR +log(Capital)+log(Labor),
            data = ex3_data,
            subset = (Sector =="Food and kindred products (20)"))
beta_20 <- fit20$coefficients[3:4]

fit36 <- lm(log(RealValueAdded)~YEAR +log(Capital)+log(Labor),
            data = ex3_data,
            subset = (Sector =="electrical and electronic machinery, equipment and supplies (36)"))
beta_36 <- fit36$coefficients[3:4]

fit37 <- lm(log(RealValueAdded)~YEAR +log(Capital)+log(Labor),
            data = ex3_data,
            subset = (Sector =="transportation equipment (37)"))
beta_37 <- fit37$coefficients[3:4]
result_beta <- matrix(c(beta_20,beta_36,beta_37),ncol = 3)
colnames(result_beta) <- c("Sector.20","Sector.36","Sector.37")
rownames(result_beta) <- c("beta1","beta2")
knitr::kable(result_beta)
```

| | Sector.20 | Sector.36 | Sector.37 |
|---|---|---|---|
| beta1 | 0.0443601 | 0.8209825 | 0.1585555 |
| beta2 | -0.9082360 | 0.8824895 | 1.1952943 |

3.

### (d)

Constraint : $\beta_1 + \beta_2 = 1$

$log(V_t) = log(\alpha) + log(\gamma)t + \beta_1 log(K_t) + (1-\beta_1)log(L_t) + log(\epsilon_t)$

$= log(\alpha) + log(\gamma)t + \beta_1 log(\frac{K_t}{L_t}) + log(L_t) + log(\epsilon_t)$

將 $YEAR$ 和 $log(\frac{K_t}{L_t})$ 當作解釋變數，以及 offset 為 $log(L_t)$。

```
fit20 <- lm(log(RealValueAdded)~YEAR +log(Capital/Labor),
            data = ex3_data,
            subset = (Sector =="Food and kindred products (20)"),offset = log(Labor))
beta_20 <- c(fit20$coefficients[3],1-fit20$coefficients[3])

fit36 <- lm(log(RealValueAdded)~YEAR +log(Capital/Labor),
            data = ex3_data,
            subset = (Sector =="electrical and electronic machinery, equipment and supplies (36)"),
            offset = log(Labor))
beta_36 <- c(fit36$coefficients[3],1-fit36$coefficients[3])

fit37 <- lm(log(RealValueAdded)~YEAR +log(Capital/Labor),
            data = ex3_data,
            subset = (Sector =="transportation equipment (37)"),offset = log(Labor))
beta_37 <- c(fit37$coefficients[3],1-fit37$coefficients[3])
result_beta <- matrix(c(beta_20,beta_36,beta_37),ncol = 3)
colnames(result_beta) <- c("Sector.20","Sector.36","Sector.37")
rownames(result_beta) <- c("beta1","beta2")
knitr::kable(result_beta)
```

| | Sector.20 | Sector.36 | Sector.37 |
|---|---|---|---|
| beta1 | -0.4947025 | 0.0345015 | -0.3168157 |
| beta2 | 1.4947025 | 0.9654985 | 1.3168157 |

Good !!!