Linear Model Assignment 2

鄭雅珊、劉奕宏、邱繼賢

Problem 1.

```
library(GGally)
```

```
dat1 <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/E2.8.txt"
                   ,header=T)
```

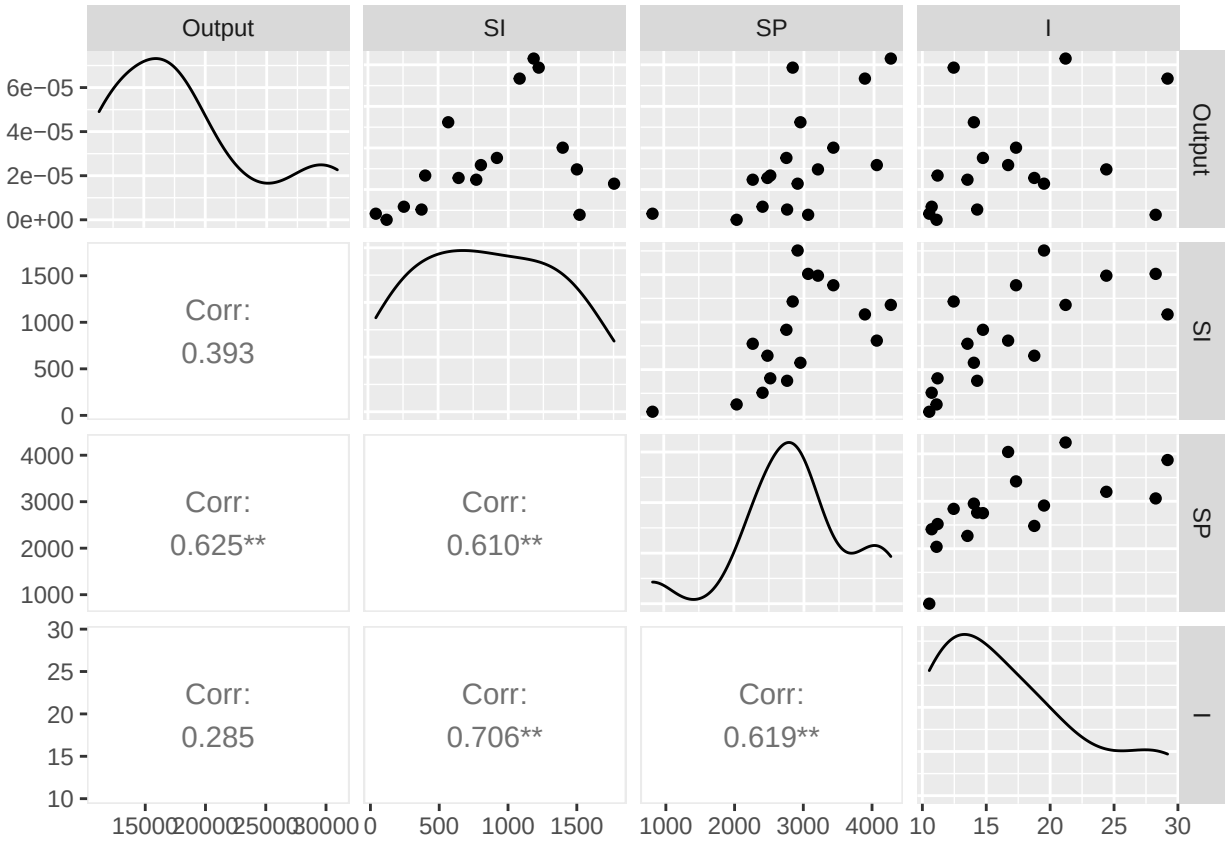
- *Output*: per capita output in Chinese yuan 人均產值 (人民幣)
- *SI*: number of workers in the factory 工廠員工數
- *SP*: land area of the factory in square meters per worker 每位員工在工廠內的使用空間 (平方公尺)
- *I*: investment in yuans per worker 每位員工的投資額

```
summary(dat1)
```

```
##      Output            SI               SP             I
##  Min.   :11360   Min.   :  56.0   Min.   : 840   Min.   :10.54
##  1st Qu.:12930   1st Qu.: 408.0   1st Qu.:2480   1st Qu.:12.45
##  Median :16680   Median : 805.0   Median :2840   Median :14.74
##  Mean   :18348   Mean   : 856.9   Mean   :2859   Mean   :16.94
##  3rd Qu.:20030   3rd Qu.:1217.0   3rd Qu.:3200   3rd Qu.:19.52
##  Max.   :30750   Max.   :1754.0   Max.   :4240   Max.   :29.19
```

- 所有變數皆為 quantitative 及 continuous，其中 *Output*、*SI* 和 *SP* 雖然皆為整數值，但因為涵蓋範圍大，可視為連續型變數。

```
ggpairs(dat1,lower = list(continuous = "cor"),
        upper = list(continuous = "points"))
```



- SI、SP 和 I 兩兩變數之間的相關係數都在 0.6 以上，顯示解釋變數間有中度以上的正相關性，從 scatter plot 也可發現此現象。
- 從 scatter plot 來看，SI 對 $Output$ 可能存在 quadratic effect，SP 和 $Output$ 看似正相關，而 I 和 $Output$ 的關係並不明確。
- SI 和 SP 的分布較對稱，$Output$ 及 I 的分布為右偏。

```
attach(dat1)
```

a.

首先配適 Model a: $Output = \beta_0 + \beta_1 \times SI + \beta_2 \times SP + \beta_3 \times I + \epsilon$

```
lm1 <- lm(Output~SI+SP+I)
summary(lm1)
```

```
##
## Call:
## lm(formula = Output ~ SI + SP + I)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6638.7 -3578.0  -558.5  4011.6  9637.4
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6026.061   5245.659   1.149   0.2713
## SI             1.742      3.777   0.461   0.6523
## SP             5.302      2.188   2.423   0.0307 *
## I           -255.506    333.194  -0.767   0.4569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5262 on 13 degrees of freedom
## Multiple R-squared:  0.4174, Adjusted R-squared:  0.2829
## F-statistic: 3.104 on 3 and 13 DF,  p-value: 0.06371
```

- $R^2 = 0.4174$ ，顯示 Output 的變異僅被模型解釋 41.74%，可能有重要的解釋變數尚未納入，配適結果只有 SP 顯著，係數為正，和前述 EDA 的圖形觀察及相關係數結果有一致，但 SI 的係數不顯著，應考慮納入二次項進行建模。
- 從 EDA 結果可知，解釋變數間的相關性高可能是造成係數估計結果不顯著的另一原因。
- $Intercept$ 及 I 的係數相較於其他變數的數量級偏大，和原始變數數量級有關，I 的數值僅介於 10-30，但 SI 的範圍落在 56-1754，SP 的範圍落在 840-4240，數量級越小的變數造成係數估計值數量級變大。

b.

考慮將 SI^2 及 $SP \times I$ 加入 Model a，接著配適

$$\text{Model b: } Output = \beta_0 + \beta_1 \times SI + \beta_2 \times SP + \beta_3 \times I + \beta_4 \times SI^2 + \beta_5 \times SP \times I + \epsilon$$

```
lm2 <- lm(Output~SI+SP+I+I(SI^2)+I(SP*I))
summary(lm2)
```

```
##
## Call:
## lm(formula = Output ~ SI + SP + I + I(SI^2) + I(SP * I))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4821.1 -1766.4  -316.4  1032.9  5638.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.240e+04  1.354e+04   3.869  0.00261 **
## SI           3.513e+01  1.029e+01   3.414  0.00579 **
## SP          -1.372e+01  4.978e+00  -2.755  0.01871 *
## I           -3.716e+03  1.001e+03  -3.710  0.00344 **
## I(SI^2)     -1.454e-02  4.894e-03  -2.971  0.01273 *
## I(SP * I)    1.022e+00  2.841e-01   3.597  0.00419 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3560 on 11 degrees of freedom
## Multiple R-squared:  0.7743, Adjusted R-squared:  0.6717
## F-statistic: 7.548 on 5 and 11 DF,  p-value: 0.002667
```

- $R^2 = 0.7743$，相較於 Model a 有大幅改進，所有變數皆顯著。
- SP 的係數相較於 Model a 正負號改變，推測是因為加入 $SP \times I$ 交互作用項，解釋了 $Output$ 的變異，進而修正 SP 對 $Output$ 的效應。
- 在 Model b 當中 I 的係數仍為負號且有顯著，顯示 I 增加會造成 $Output$ 減少，另外 SP 增加也會造成 $Output$ 減少，但是 I 和 SP 兩個變數間存在交互作用，當兩者同時增加或同時減少時，可以使得 $Output$ 增加。

c.

倘若把 Model b 當作真實模型進行推論，可寫出目標函數

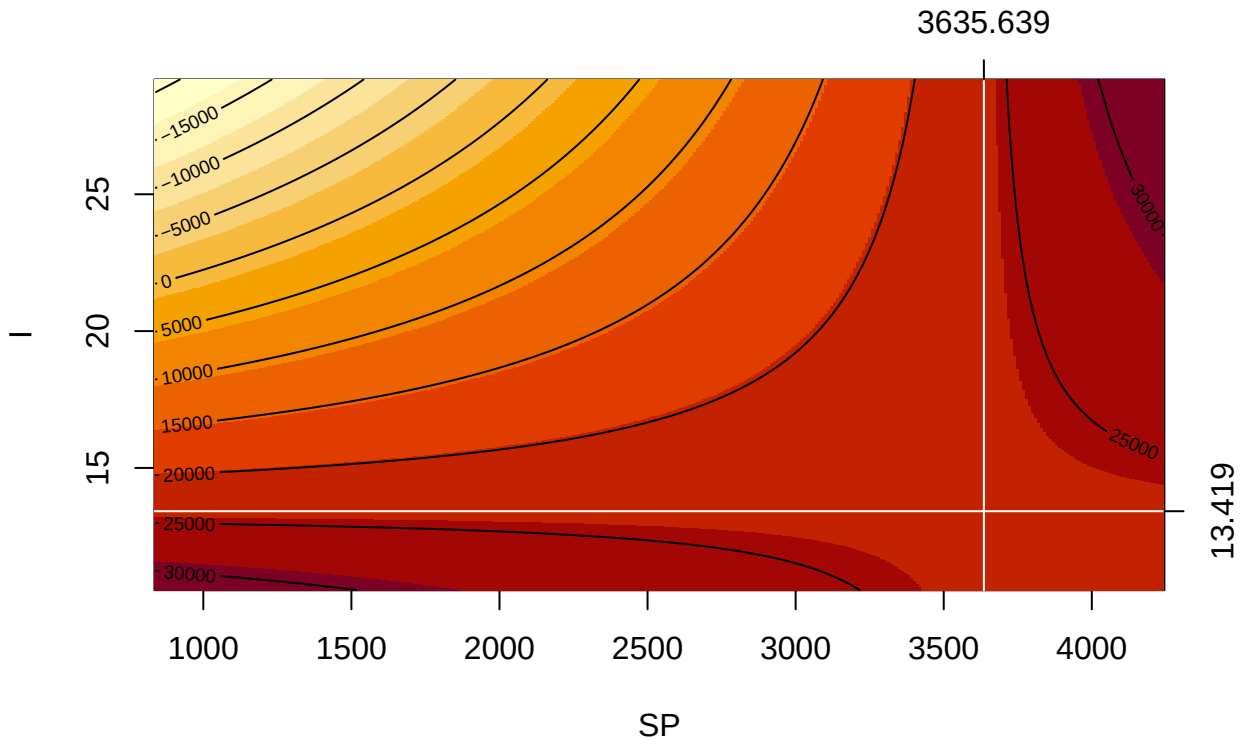
$$f(SI, SP, I) = 52400 + 35.13 \times SI - 0.01454 \times SI^2 - 13.72S \times P - 3716 \times I + 1.022 \times SP \times I$$

- f 對 SI 作一階偏微分可得 $\partial f/\partial SI = 35.13 - 2*0.01454*SI = 0 \Rightarrow SI = 35.13/(2*0.01454) = 1208.221$。且二階偏微分為 $\partial^2 f/\partial SI^2 = -2*0.01454 < 0$。因此固定 SP 及 I 的值，當 SI 為 1208.221 時 $Output$ 最大。由於 SI 為員工數，應為整數值較合理，因此比較靠近最佳值的兩個整數點 $SI = 1208$ 及 $SI = 1209$ 代入目標函數，發現當 $SI = 1208$ 時的目標函數值較大，因此以下用 $SI = 1208$ 接續討論。
- f 對 SP 作一階偏微分可得 $\partial f/\partial SP = -13.72 + 1.022 * I = 0 \Rightarrow I = 13.72/1.022 = 13.419$。當 I 大於 13.419 時，$Output$ 隨著 SP 增加而遞增；反之，$Output$ 隨著 SP 增加而遞減。
- f 對 I 作一階偏微分可得 $\partial f/\partial I = -3716 + 1.022 * SP = 0 \Rightarrow SP = 3716/1.022 = 3635.639$。當 SP 大於 3635.639 時，$Output$ 隨著 I 增加而遞增；反之，$Output$ 隨著 I 增加而遞減。

```
SI_o <- round(-lm2$coefficients[2]/(2*lm2$coefficients[5]))
SP_o <- -lm2$coefficients[4]/lm2$coefficients[6]
I_o <- -lm2$coefficients[3]/lm2$coefficients[6]

output_fun <- function(par){
  sp <- par[1]
  i <- par[2]
  -t(c(1,SI_o,sp,i,SI_o^2,sp*i))%*%lm2$coefficients
}
SP_seq <- seq(min(SP),max(SP),10)
I_seq <- seq(min(I),max(I),0.01)
SP_v <- rep(SP_seq,each=length(I_seq))
I_v <- rep(I_seq,length(SP_seq))
out <- -apply(cbind(SP_v,I_v),1,output_fun)

image(SP_seq, I_seq, matrix(out,length(SP_seq), length(I_seq), byrow=T),
      xlab="SP", ylab="I")
contour(SP_seq, I_seq, matrix(out,length(SP_seq), length(I_seq), byrow=T),add=T)
abline(h=I_o,v=SP_o,col="white")
axis(4,at=I_o,round(I_o,3))
axis(3,at=SP_o,round(SP_o,3))
```



- 若進一步從 contour plot 來看，圖中等高線數值表示 *Output* 的值，水平白線表示 $I = 13.419$，垂直白線為 $SP = 3635.639$，為 *Output* 條件遞增或遞減的分界點。*Output* 沿著左上至右下的方向遞增，但通過臨界範圍 $(I = 13.419, SP = 3635.639)$ 後則轉往 I 和 SP 同時增加或同時遞減的方向遞增，也可以看出 Output 最大值會落在左下角 (SP 及 I 值皆小) 或右上角 (SP 及 I 值皆大) 的位置。
- 承上述可知，若 I 增加但 SP 減少，反而會使得 *Output* 下降，甚至出現負值的情形。
- 為了預測上的正確性，此處在資料範圍內討論 SP 及 I 的最佳解。

```
optim(c(mean(SP),mean(I)),output_fun,method="L-BFGS-B",lower=c(min(SP),min(I)),
      upper=c(max(SP),max(I)))
```

```
## $par
## [1] 840.00  10.54
##
## $value
## [1] -31990.8
##
## $counts
## function gradient
##        9        9
##
## $convergence
## [1] 0
##
## $message
```

```
## [1] "CONVERGENCE: NORM OF PROJECTED GRADIENT <= PGTOL"
```

```
optim(c(max(SP),min(I)),output_fun,method="L-BFGS-B",lower=c(min(SP),min(I)),
      upper=c(max(SP),max(I)))
```

```
## $par
## [1] 4240.00   29.19
##
## $value
## [1] -33506.56
##
## $counts
## function gradient
##        4        4
##
## $convergence
## [1] 0
##
## $message
## [1] "CONVERGENCE: NORM OF PROJECTED GRADIENT <= PGTOL"
```

```
t(c(1,SI_o,min(SP),min(I),SI_o^2,min(SP)*min(I)))%*%lm2$coefficients
```

```
##         [,1]
## [1,] 31990.8
```

```
t(c(1,SI_o,max(SP),max(I),SI_o^2,max(SP)*max(I)))%*%lm2$coefficients
```

```
##          [,1]
## [1,] 33506.56
```

- 在固定 $SI = 1208$，若從不同的起始值搜尋，在圖形中左下角找到的是當 $SP = 840, I = 10.54$，最大值 $Output = 31990.8$，而在圖形的右上方找到的是當 $SP = 4240, I = 29.19$，最大值 $Output = 33506.56$。
- 從結果進行解釋，增加員工的投資額及縮減工作空間，會造成人均產值減少，相反地，降低人事成本與擴充工作空間可有效提升人均產值，另一方面，當人事成本低於某一金額 (此例為 $I = 13.419$) ，此時再減少工作空間，可再提升人均產值，而當工作空間大於特定占地面積 (此例為 $SP = 3635.639$)，增加員工投資額可再提升人均產值，上述兩種情況可能對應到特定的產業或工作項目。
- 此外，一間工廠的員工數，太多或太少皆無益於人均產值的提升，以這筆資料來看，最佳解 $SI = 1208$ 約落在靠近 $Q3$ 的位置。
- 站在資方的立場，當實際產值差異不大時，考慮投資成本較低的方案是較符合經濟效益的。

Problem 2.

```
library(dplyr)
library(knitr)
library(ggplot2)
```

```
prostate <- read.table("prostate.txt")
```

a.

這裡我們把 **lpsa** 當作 response variable、**lcavol** 當作 predictor，fit 一個 Linear model。

$$\text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \varepsilon$$

```
fit1 <- lm(lpsa ~ lcavol, data = prostate)
summary(fit1)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67625 -0.41648  0.09859  0.50709  1.89673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50730    0.12194   12.36   <2e-16 ***
## lcavol       0.71932    0.06819   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7875 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

從 summary 中可以看到 Residual standard error($\hat{\sigma}$) $= 0.7875$，$R^2 = 0.5394$。

b.

現在我們把 lweight, svi, lbph, age, lcp, pgg45, gleason 一個一個依序加進我們的 model 裡面並且記錄下這 7 個 model 的 $\hat{\sigma}$ 和 R^2，並且觀察他們的變化。

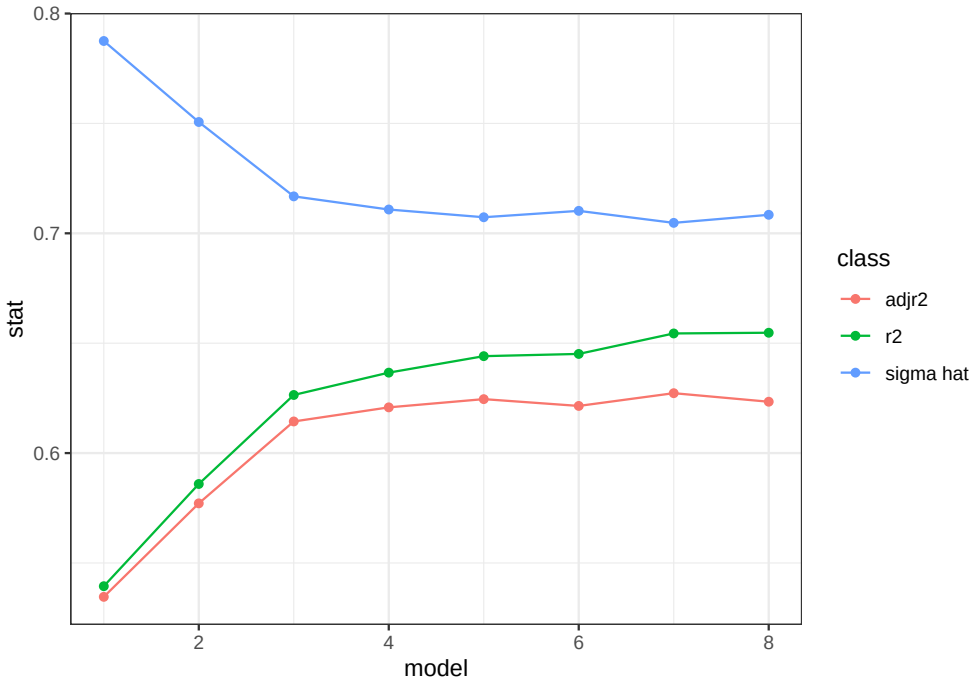
$$
\begin{aligned}
\text{model 1} &: \text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \varepsilon \\
\text{model 2} &: \text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \beta_2 \times \text{lweight} + \varepsilon \\
\text{model 3} &: \text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \beta_2 \times \text{lweight} + \beta_3 \times \text{svi} + \varepsilon \\
\text{model 4} &: \text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \beta_2 \times \text{lweight} + \beta_3 \times \text{svi} + \beta_4 \times \text{lbph} + \varepsilon \\
\text{model 5} &: \text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \beta_2 \times \text{lweight} + \beta_3 \times \text{svi} + \beta_4 \times \text{lbph} + \beta_5 \times \text{age} + \varepsilon \\
\text{model 6} &: \text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \beta_2 \times \text{lweight} + \beta_3 \times \text{svi} + \beta_4 \times \text{lbph} + \beta_5 \times \text{age} + \beta_6 \times \text{lcp} + \varepsilon \\
\text{model 7} &: \text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \beta_2 \times \text{lweight} + \beta_3 \times \text{svi} + \beta_4 \times \text{lbph} + \beta_5 \times \text{age} + \beta_6 \times \text{lcp} + \beta_7 \times \text{pgg45} + \varepsilon \\
\text{model 8} &: \text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \beta_2 \times \text{lweight} + \beta_3 \times \text{svi} + \beta_4 \times \text{lbph} + \beta_5 \times \text{age} + \beta_6 \times \text{lcp} + \beta_7 \times \text{pgg45} \\
&\quad + \beta_8 \times \text{gleason} + \varepsilon
\end{aligned}
$$

```
summary_1 <- summary(fit1)
summary_2 <- lm(lpsa ~ lcavol + lweight, data = prostate) %>% summary()
summary_3 <- lm(lpsa ~ lcavol + lweight + svi, data = prostate) %>% summary()
summary_4 <- lm(lpsa ~ lcavol + lweight + svi + lbph, data = prostate) %>% summary()
summary_5 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age,
                    data = prostate) %>% summary()
summary_6 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp,
                     data = prostate) %>% summary()
summary_7 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp + pgg45,
                      data = prostate) %>% summary()
summary_8 <- lm(lpsa ~ ., data = prostate) %>% summary()
total_mse <- c(summary_1$sigma, summary_2$sigma, summary_3$sigma, summary_4$sigma,
               summary_5$sigma, summary_6$sigma, summary_7$sigma,
               summary_8$sigma)
total_r2 <- c(summary_1$r.squared, summary_2$r.squared, summary_3$r.squared,
              summary_4$r.squared, summary_5$r.squared, summary_6$r.squared,
              summary_7$r.squared, summary_8$r.squared)
total_adjr2 <- c(summary_1$adj.r.squared, summary_2$adj.r.squared,
                 summary_3$adj.r.squared, summary_4$adj.r.squared,
                 summary_5$adj.r.squared, summary_6$adj.r.squared,
                 summary_7$adj.r.squared, summary_8$adj.r.squared)
```

```
matrix(round(c(total_mse, total_r2, total_adjr2),3), ncol = 3) %>%
  `colnames<-`(c('$\\hat{\\sigma}$', '$R^2$', '$R_{adj}^2$')) %>%
  `rownames<-`(c('model 1', 'model 2', 'model 3', 'model 4', 'model 5',
                 'model 6', 'model 7', 'model 8')) %>% kable()
```

	$\hat{\sigma}$	R^2	R^2_{adj}
model 1	0.787	0.539	0.535
model 2	0.751	0.586	0.577
model 3	0.717	0.626	0.614
model 4	0.711	0.637	0.621
model 5	0.707	0.644	0.625
model 6	0.710	0.645	0.621
model 7	0.705	0.654	0.627
model 8	0.708	0.655	0.623

```
statistic <- data.frame(stat=c(total_mse, total_r2, total_adjr2),
                        class=rep(c('sigma hat','r2','adjr2'), each=8),
                        model = rep(1:8, 3))
ggplot(data = statistic, aes(x = model, y = stat, col = class))+
  geom_point()+
  geom_line()+
  theme_bw()
```



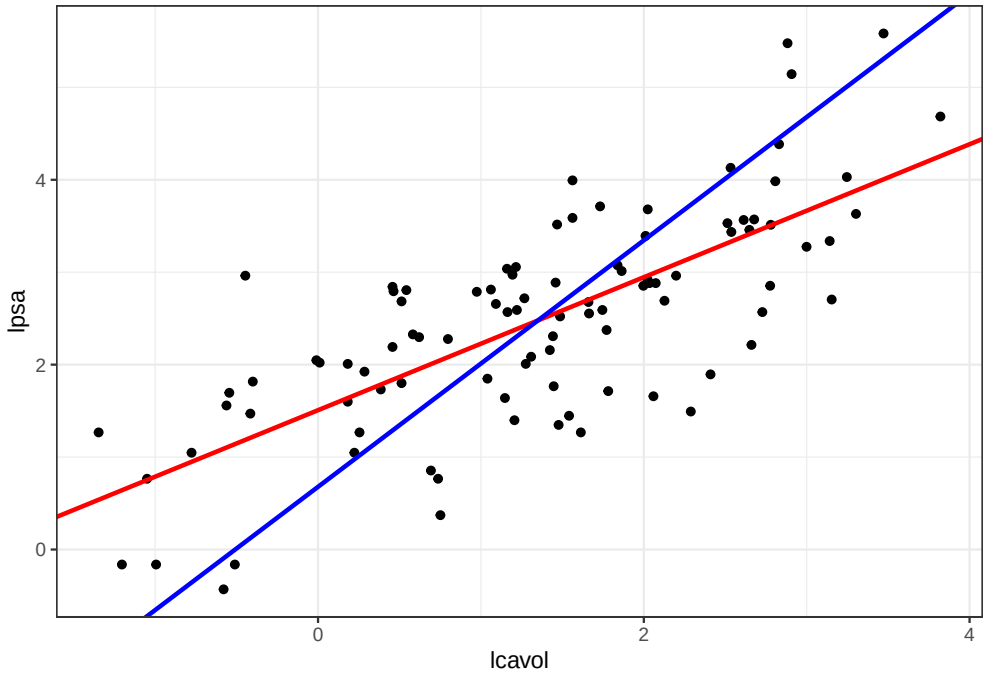
從上面的 table 和圖我們可以看到，由於 model 1~8 中任兩個相鄰的 model，上一個 model 接有被包含在下一個 model 裡，所以導致 R^2 會持續上升，但是 R^2_{adj} 因為有 degree of freedom 的校正，所以當加入比較不顯著的變數的時候並不會由於變數的增加而上升；同理因為 model 間的包含關係，RSS 也會持續下降，但由於 residual standard error($\hat{\sigma} = \sqrt{\frac{\text{RSS}}{n-p}}$) 是有除過 degree of freedom 的指標，所以結果會跟 R^2_{adj} 類似，當加入比較不顯著的變數的時候一樣不會由於變數的加入而減少。

c.

我們先 fit 兩個 model

$$\text{model 9} \ : \ \text{lpsa} = \beta_0 + \beta_1 \times \text{lcavol} + \varepsilon$$
$$\text{model 10} \ : \ \text{lcavol} = \beta_0 + \beta_1 \times \text{lpsa} + \varepsilon$$

```
fit10 <- lm(lcavol ~ lpsa, data = prostate)
ggplot(data = prostate, aes(x = lcavol, y = lpsa))+
  geom_point()+
  theme_bw()+
  geom_abline(intercept = fit1$coefficients[1], slope = fit1$coefficients[2],
              col = "red", lwd = 1)+
  geom_abline(intercept = -fit10$coefficients[1]/fit10$coefficients[2],
              slope = 1/fit10$coefficients[2], col = "blue", lwd = 1)
```




```
matrix(c(fit1$coefficients, fit10$coefficients), ncol = 2, byrow = T) %>%
  `colnames<-`(c('$\\hat{\\beta_0}$', '$\\hat{\\beta_1}$')) %>%
  `rownames<-`(c("model 9", "model 10")) %>%
  kable()
```

	$\hat{\beta}_0$	$\hat{\beta}_1$
model 9	1.5072979	0.7193201
model 10	-0.5085802	0.7499191

從上圖可以看到，紅色的線代表的是 $lpsa \sim lcavol$，藍色的線代表的是 $lcavol \sim lpsa$。而從 table 中我們可以得到

$$\begin{aligned}\text{model 9} &: \text{lpsa} = 1.5072979 + 0.7193201 \times \text{lcavol} \\ \text{model 10} &: \text{lcavol} = -0.5085802 + 0.7499191 \times \text{lpsa} \\ &\Rightarrow \text{lpsa} = -\frac{(-0.5085802)}{0.7499191} + \frac{1}{0.7499191} \times \text{lcavol}\end{aligned}$$

```
matrix(c(mean(prostate$lcavol), mean(prostate$lpsa)), nrow = 2) %>%
  `colnames<-`('Mean') %>% `rownames<-`(c('lcavol', 'lpsa')) %>% kable()
```

	Mean
lcavol	1.350010
lpsa	2.478387

由於我們知道在 simple linear regression 中，(mean of response, mean of predictor) 這個點會通過回歸線，所以我們可以得知上述兩個 model 的交點就是 $(\overline{lcavol}, \overline{lpsa}) = (1.35001, 2.478387)$。

Problem 3.

將資料根據不同的 economic sector 分成三份：

```
data3 = read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/E2.9.txt",
                   skip = 2)
data3_20 = data3[,c(1,2,5,8)] ; colnames(data3_20) = c("year","capital","labor","RVA")
data3_36 = data3[,c(1,3,6,9)] ; colnames(data3_36) = c("year","capital","labor","RVA")
data3_37 = data3[,c(1,4,7,10)] ; colnames(data3_37) = c("year","capital","labor","RVA")
```

a.

對三份資料分別建構模型：

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

response variable : $\log(V_t)$
predictor variable : $\log(K_t)$, $\log(L_t)$
error term : $\log(\epsilon_t)$ with zero mean and constant variance

```
model_3a_20 = lm(log(RVA) ~ I(log(capital)) + I(log(labor)), data3_20)
model_3a_36 = lm(log(RVA) ~ I(log(capital)) + I(log(labor)), data3_36)
model_3a_37 = lm(log(RVA) ~ I(log(capital)) + I(log(labor)), data3_37)
```

對三模型分別估計參數 $\hat{\beta}_1$, $\hat{\beta}_2$ 呈現如下表：

```
library(knitr)
rname = c("Food and kindred products (20)",
          "Equipment and supplies (36)",
          "Transportation equipment (37)")
coef_table = rbind(model_3a_20$coef[-1],model_3a_36$coef[-1],model_3a_37$coef[-1])
rownames(coef_table) = rname
kable(coef_table, col.names = c("$\\hat{\\beta_1}$", "$\\hat{\\beta_2}$"),digits = 3)
```

	$\hat{\beta}_1$	$\hat{\beta}_2$
Food and kindred products (20)	0.227	-1.458
Equipment and supplies (36)	0.526	0.254
Transportation equipment (37)	0.506	0.845

b.

加上條件 $\beta_1 + \beta_2 = 1$，改寫模型：

$$\begin{aligned}\log(V_t) &= \log(\alpha) + \beta_1 \log(K_t) + (1-\beta_1)\log(L_t) + \log(\epsilon_t) \\ &= \log(L_t) + \log(\alpha) + \beta_1 \log\left(\frac{K_t}{L_t}\right) + \log(\epsilon_t)\end{aligned}$$

response variable : $\log(V_t)$
predictor variable : $\log\left(\frac{K_t}{L_t}\right)$
offset : $\log(L_t)$
error term : $\log(\epsilon_t)$ with zero mean and constant variance

```
model_3b_20 = lm(log(RVA) ~ offset(log(labor)) + I(log(capital/labor)), data3_20)
model_3b_36 = lm(log(RVA) ~ offset(log(labor)) + I(log(capital/labor)), data3_36)
model_3b_37 = lm(log(RVA) ~ offset(log(labor)) + I(log(capital/labor)), data3_37)
```

對三模型分別估計參數 $\hat{\beta}_1$, $\hat{\beta}_2 = 1 - \hat{\beta}_1$ 呈現如下表：

```
coef_20 = c(model_3b_20$coef[2],1-model_3b_20$coef[2])
coef_36 = c(model_3b_36$coef[2],1-model_3b_36$coef[2])
coef_37 = c(model_3b_37$coef[2],1-model_3b_37$coef[2])
coef_table = rbind(coef_20,coef_36,coef_37)
rownames(coef_table) = rname
kable(coef_table, col.names = c("$\\hat{\\beta_1}$", "$\\hat{\\beta_2}$"),digits = 3)
```

	$\hat{\beta}_1$	$\hat{\beta}_2$
Food and kindred products (20)	1.29	-0.29
Equipment and supplies (36)	0.90	0.10
Transportation equipment (37)	0.01	0.99

c.

多考慮 technological development γ^t，這是一個跟年份 t 有關的變化，分別對三份資料建構模型：

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

response variable : $\log(V_t)$
predictor variable : t , $\log(K_t)$, $\log(L_t)$
error term : $\log(\epsilon_t)$ with zero mean and constant variance

```
model_3c_20 = lm(log(RVA) ~ year + I(log(capital)) + I(log(labor)), data3_20)
model_3c_36 = lm(log(RVA) ~ year + I(log(capital)) + I(log(labor)), data3_36)
model_3c_37 = lm(log(RVA) ~ year + I(log(capital)) + I(log(labor)), data3_37)
```

對三模型分別估計參數 $\hat{\beta}_1$, $\hat{\beta}_2$ 呈現如下表：

```
coef_table = rbind(model_3c_20$coef[3:4],model_3c_36$coef[3:4],model_3c_37$coef[3:4])
rownames(coef_table) = rname
kable(coef_table, col.names = c("$\\hat{\\beta_1}$", "$\\hat{\\beta_2}$"),digits = 3)
```

	$\hat{\beta}_1$	$\hat{\beta}_2$
Food and kindred products (20)	0.044	-0.908
Equipment and supplies (36)	0.821	0.882
Transportation equipment (37)	0.159	1.195

d.

加上條件 $\beta_1 + \beta_2 = 1$，改寫模型：

$$\begin{aligned}\log(V_t) &= \log(\alpha) + t\log(\gamma) + \beta_1 \log(K_t) + (1-\beta_1)\log(L_t) + \log(\epsilon_t) \\ &= \log(L_t) + \log(\alpha) + t\log(\gamma) + \beta_1 \log\left(\frac{K_t}{L_t}\right) + \log(\epsilon_t)\end{aligned}$$

response variable : $\log(V_t)$
predictor variable : t , $\log\left(\frac{K_t}{L_t}\right)$
offset : $\log(L_t)$
error term : $\log(\epsilon_t)$ with zero mean and constant variance

```
model_3d_20 = lm(log(RVA) ~ offset(log(labor)) + year + I(log(capital/labor)), data3_20)
model_3d_36 = lm(log(RVA) ~ offset(log(labor)) + year + I(log(capital/labor)), data3_36)
model_3d_37 = lm(log(RVA) ~ offset(log(labor)) + year + I(log(capital/labor)), data3_37)
```

對三模型分別估計參數 $\hat{\beta}_1$, $\hat{\beta}_2 = 1 - \hat{\beta}_1$ 呈現如下表：

```
coef_20 = c(model_3d_20$coef[3],1-model_3d_20$coef[3])
coef_36 = c(model_3d_36$coef[3],1-model_3d_36$coef[3])
coef_37 = c(model_3d_37$coef[3],1-model_3d_37$coef[3])
coef_table = rbind(coef_20,coef_36,coef_37)
rownames(coef_table) = rname
kable(coef_table, col.names = c("$\\hat{\\beta_1}$", "$\\hat{\\beta_2}$"),digits = 3)
```

	$\hat{\beta}_1$	$\hat{\beta}_2$
Food and kindred products (20)	-0.495	1.495
Equipment and supplies (36)	0.035	0.965
Transportation equipment (37)	-0.317	1.317