# Linear Model Assignment 3

110024516 統研碩一邱繼賢

2021 年 11 月 3 日

**1.a.**
**i.**
Full model :

$$press \ = \ \beta_0 + \beta_1 \ HCHO + \beta_2 \ catalyst + \beta_3 \ temp + \beta_4 \ time + \epsilon$$

$$\begin{cases} H_0 \ : \ \beta_1 = \beta_1 = \beta_3 = \beta_4 = 0 \\ H_1 \ : \ at \ least \ one \ \beta_i \ \neq \ 0, \ i \ = 1,2,3,4 \end{cases}$$

```
df = read.table("wrinkle.txt", header = T)
model_a = lm(press ~ HCHO + catalyst + temp + time, data = df)
summary(model_a)
```

```
##
## Call:
## lm(formula = press ~ HCHO + catalyst + temp + time, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07876 -0.63939 -0.08531  0.36236  1.65332
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.912212   0.875484  -1.042   0.3074
## HCHO         0.160726   0.066166   2.429   0.0227 *
## catalyst     0.219783   0.034062   6.452 9.33e-07 ***
## temp         0.011226   0.004973   2.257   0.0330 *
## time         0.101974   0.058735   1.736   0.0948 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8365 on 25 degrees of freedom
## Multiple R-squared:  0.6924, Adjusted R-squared:  0.6432
## F-statistic: 14.07 on 4 and 25 DF,  p-value: 3.845e-06
```

$\therefore$ the test statistic $F \ = \ 14.07$, and the $p-value \ = \ 3.845 \ \times \ 10^{-6} \ < \ \alpha \ = \ 0.05$

$\Rightarrow \ reject \ H_0$.

Thus at least one of the 4 predictors are significant.

**ii.** Use model 2 :

$$press \ = \ \beta_0 + \beta_1 \ HCHO + \beta_2 \ catalyst + \beta_3 \ temp + \epsilon$$

to compare with the full model as above.

$$\begin{cases} H_0 & : \ \beta_4 \ = \ 0 \ (model\ 2) \\ H_1 & : \ \beta_4 \ \neq \ 0 \ (full\ model) \end{cases}$$

```
model_2 = lm(press ~ HCHO + catalyst + temp, data = df)
anova(model_2, model_a)
```

```
## Analysis of Variance Table
##
## Model 1: press ~ HCHO + catalyst + temp
## Model 2: press ~ HCHO + catalyst + temp + time
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     26 19.605
## 2     25 17.495  1    2.1094 3.0143 0.09484 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\therefore$ the test statistic $F \ = \ 3.0143$, and the $p-value \ = \ 0.09484 \ > \ \alpha \ = \ 0.05$

$\Rightarrow \ fail\ to\ reject\ H_0$.

Thus, we do not have enough evidence to show that the predictor *time* is significant when the predictors $HCHO$, *catalyst*, *temp* are in the model.

**iii.** Model 3 :

$$press \ = \ \beta_0 + \beta_4 \ time + \epsilon$$

$$\begin{cases} H_0 & : \ \beta_4 \ = \ 0 \\ H_1 & : \ \beta_4 \ \neq \ 0 \end{cases}$$

```
model_3 = lm(press ~ time, data = df)
summary(model_3)
```

```
##
## Call:
## lm(formula = press ~ time, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3108 -1.5857  0.9376  1.1842  1.2876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.36665    0.46735   7.204 7.69e-08 ***
## time         0.04916    0.09889   0.497    0.623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.419 on 28 degrees of freedom
## Multiple R-squared:  0.008747,  Adjusted R-squared:  -0.02665
## F-statistic: 0.2471 on 1 and 28 DF,  p-value: 0.623
```

$\therefore$ the test statistic $F = 0.2471$, $t = 0.497$ where $F = t^2$, and the $p-value = 0.623 > \alpha = 0.05$

$\Rightarrow$ *fail to reject* $H_0$.

Thus, we do not have enough evidence to show that the predictor *time* is significant.

Compares to problem ii., the $p-value$ in this problem(namely 0.623) is bigger than the $p-value$ in the above problem(namely 0.09484).

如果各變數之間具有直交性，則兩題所計算出的 $p-value$ 會一樣，但此題明顯無此現象。

**iv.** Use model 4 :

$$press = \beta_0 + \beta_1^{\star}\ (HCHO - catalyst) + \beta_3\ temp + \beta_4\ time + \epsilon$$

to compare with the full model.

$$\begin{cases} H_0 : \beta_1 = -\beta_2 = \beta_1^{\star}\ (model\ 4) \\ H_1 : \beta_1 \neq -\beta_2\ (full\ model) \end{cases}$$

```
model_4 = lm(press ~ I(HCHO-catalyst) + temp + time, data = df)
anova(model_4, model_a)
```

```
## Analysis of Variance Table
##
## Model 1: press ~ I(HCHO - catalyst) + temp + time
## Model 2: press ~ HCHO + catalyst + temp + time
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     26 36.449
## 2     25 17.495  1    18.954 27.085 2.199e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\therefore$ the test statistic $F = 27.085$, and the $p-value = 2.199 \times 10^{-5} < \alpha = 0.05$

$\Rightarrow$ *reject* $H_0$.

Thus, we have enough evidence to show that $\beta_1 \neq -\beta_2$. Equivalently, it shows that there is evidence that $HCHO$ and $catalyst$ need to be treated separately instead of being treated as $(HCHO - catalyst)$ in the context of this particular model.

**v.** Use model 5 :

$$press = \beta_0 + 0.25\ HCHO + \beta_2\ catalyst + \beta_3\ temp + \beta_4\ time + \epsilon$$

to compare with the full model.

$$\begin{cases} H_0 : \beta_1 = 0.25\ (model\ 5) \\ H_1 : \beta_1 \neq 0.25\ (full\ model) \end{cases}$$

```
model_5 = lm(press ~ catalyst + temp + time + offset(0.25*HCHO), data = df)
anova(model_5, model_a)
```

```
## Analysis of Variance Table
##
## Model 1: press ~ catalyst + temp + time + offset(0.25 * HCHO)
## Model 2: press ~ HCHO + catalyst + temp + time
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 18.769
## 2     25 17.495  1    1.2739 1.8204 0.1894
```

$\therefore$ the test statistic $F \ = \ 1.8204$, and the $p - value \ = \ 0.1894 \ > \ \alpha \ = \ 0.05$

$\Rightarrow$ *fail to reject* $H_0$.

Thus, we do not have enough evidence to show that the regression parameter associated with $HCHO$ namely $\beta_1 \ \neq \ 0.25$, when the predictors *catalyst*, *temp*, *time* are in the model.

**vi.** Use model 6 :

$$press \ = \ \beta_1 \ HCHO + \beta_2 \ catalyst + \beta_3 \ temp + \beta_4 \ time + \beta_5 \ temp^2 + \beta_6 \ time^2 + \beta_7(temp \ \times \ time) + \epsilon$$

to compare the full model.

$$\begin{cases} H_0 \ : \ model \ 6 \ fits \ better \\ H_1 \ : \ full \ model \ fits \ better \end{cases}$$

```
model_6 = lm(press ~ HCHO + catalyst + temp + time +
                 I(temp^2) + I(time^2) + I(temp*time),
             data = df)
anova(model_6, model_a)
```

```
## Analysis of Variance Table
##
## Model 1: press ~ HCHO + catalyst + temp + time + I(temp^2) + I(time^2) +
##     I(temp * time)
## Model 2: press ~ HCHO + catalyst + temp + time
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     22 12.677
## 2     25 17.495 -3    -4.818 2.7871 0.06462 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\therefore$ the test statistic $F \ = \ 2.7871$, and the $p - value \ = \ 0.06462 \ > \ \alpha \ = \ 0.05$

$\Rightarrow$ *fail to reject* $H_0$.

Thus, we do not have enough evidence to show that full model fits better than model 6.

**b.** Model b :

$$log(5 - press) \ = \ \alpha_0 + \alpha_1 \ HCHO + \alpha_2 \ catalyst + \alpha_3 \ temp + \alpha_4 \ time + \delta$$

$$\begin{cases} H_0 \ : \ \alpha_1 = \alpha_1 = \alpha_3 = \alpha_4 = 0 \\ H_1 \ : \ at \ least \ one \ \alpha_i \ \neq \ 0, \ i \ = 1, 2, 3, 4 \end{cases}$$

```r
model_b = lm(log(5-press) ~ HCHO + catalyst + temp + time, data = df)
summary(model_b)
```

```
##
## Call:
## lm(formula = log(5 - press) ~ HCHO + catalyst + temp + time,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8405 -0.5311  0.1705  0.5152  1.0916
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.302298   0.806402   4.095 0.000388 ***
## HCHO        -0.197622   0.060945  -3.243 0.003347 **
## catalyst    -0.169486   0.031375  -5.402 1.32e-05 ***
## temp        -0.005848   0.004580  -1.277 0.213425
## time        -0.091855   0.054100  -1.698 0.101951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7705 on 25 degrees of freedom
## Multiple R-squared:  0.6418, Adjusted R-squared:  0.5845
## F-statistic:  11.2 on 4 and 25 DF,  p-value: 2.406e-05
```

(1) 使用 $press$ 作為反應變數的模型 (以下簡稱 model a) 和使用 $log(5 - press)$ 作為反應變數的模型 (以下簡稱 model b)，在 overall test 中皆呈現為顯著，但在各單項變數的檢定中就有不同，model a 對 $HCHO$, $catalyst$, $temp$ 三個變數結果皆呈現顯著，但 model b 只對 $HCHO$, $catalyst$ 兩變數結果呈現為顯著。

(2) $R^2$ 和 $Adj - R^2$ 兩模型呈現結果數值差異不大，但都是 model a 的偏大。

(3) $Residual \ standard \ error$ 則是 model b 的數值比較小。

**c.**

Model c1($\Omega_1$) : $press \ = \ \beta_0 + \beta_1 \ HCHO + \epsilon$ , ($\omega_1$) : $press \ = \ \beta_0 + \epsilon$
Model c2($\Omega_2$) : $HCHO \ = \ \alpha_0 + \alpha_1 \ press + \delta$ , ($\omega_2$) : $HCHO \ = \ \alpha_0 + \delta$

$\Rightarrow \ \beta_1 \ = \ \frac{1}{\alpha_1}$

```r
model_c1 = lm(press ~ HCHO, data = df)
model_c2 = lm(HCHO ~ press, data = df)
summary(model_c1)
```

```
##
## Call:
## lm(formula = press ~ HCHO, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5253 -1.3362  0.6358  1.0888  1.6304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.4139     0.6897   3.500  0.00158 **
## HCHO          0.1889     0.1062   1.779  0.08605 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.351 on 28 degrees of freedom
## Multiple R-squared:  0.1016, Adjusted R-squared:  0.0695
## F-statistic: 3.166 on 1 and 28 DF,  p-value: 0.08605
```

```r
summary(model_c2)
```

```
##
## Call:
## lm(formula = HCHO ~ press, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8193 -1.5721  0.2076  1.3607  4.4495
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.1525     1.1535   3.600  0.00121 **
## press         0.5377     0.3022   1.779  0.08605 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.279 on 28 degrees of freedom
## Multiple R-squared:  0.1016, Adjusted R-squared:  0.0695
## F-statistic: 3.166 on 1 and 28 DF,  p-value: 0.08605
```

$$Note\ that\ \ Y\ :\ press,\ X\ :\ HCHO$$

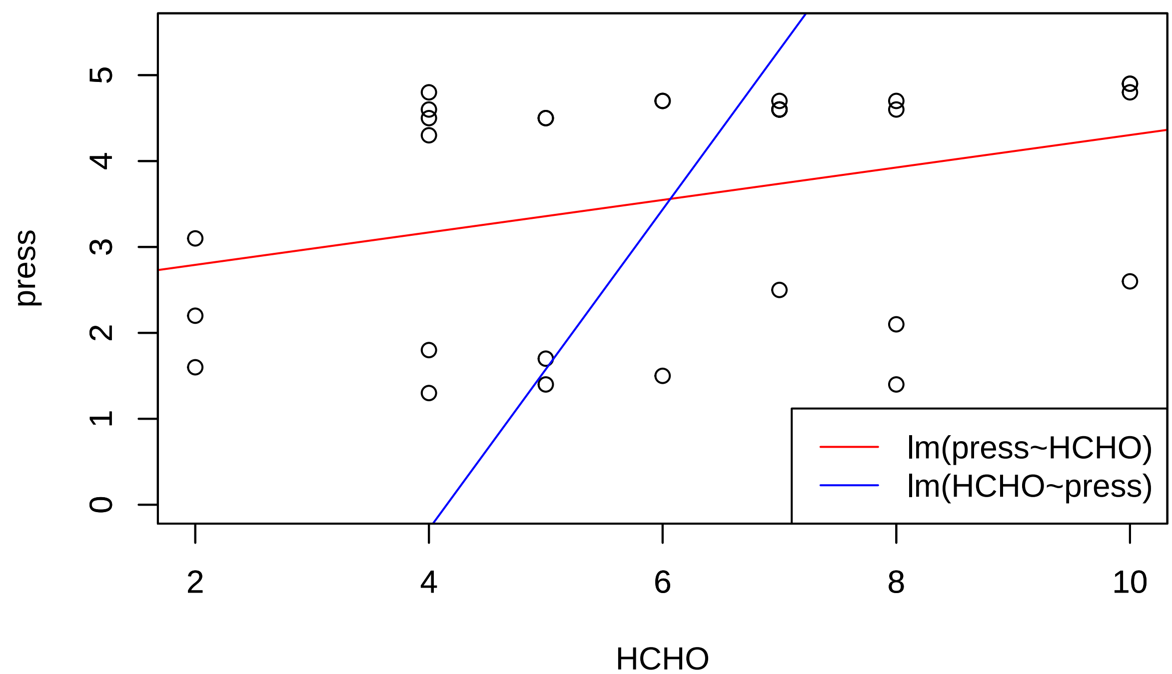
$$R_1^2 \ = \ 1 - \frac{RSS_{\Omega_1}}{TSS_{\omega_1}} \ = \ (\frac{\sum(y_i - \overline{y})(\hat{y}_i - \overline{y})}{\sqrt{\sum(y_i - \overline{y})^2 \sum(\hat{y}_i - \overline{y})^2}})^2 \ = \ (cor(Y,\ \hat{Y}))^2 \ = \ (cor(Y,\ \hat{\beta}_1 X))^2$$

$$= \ (cor(Y,\ \frac{1}{\hat{\alpha}_1} X))^2 \ = \ (cor(\hat{\alpha}_1 Y,\ X))^2 \ = \ (cor(X,\ \hat{X}))^2 \ = \ R_2^2$$

$\Rightarrow$ The $R^2$ of model c1 and model c2 are the same, and $F \ = \ \frac{R^2}{1-R^2}\frac{(n-p)}{(p-q)}$

$\therefore$ The test statistic $F$ and the $p-value$ of the two models are all the same.

```r
plot(df$HCHO, df$press, xlab = "HCHO", ylab = "press", ylim = c(0, 5.5))
abline(model_c1, col = "red")
slope2 = 1/model_c2$coefficients[2]
intercept2 = -model_c2$coefficients[1]*slope2
par(new = T)
curve(slope2*x+intercept2, 0,11, xlim = c(2,10), ylim = c(0, 5.5),
      xlab = "", ylab = "", col = "blue")
legend( x = "bottomright",
        legend = c("lm(press~HCHO)","lm(HCHO~press)"),
        col = c("red","blue"), lwd = 1, lty = c(1,1), merge = FALSE)
```




$$
\begin{aligned}
\hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2} \\
\hat{\alpha}_1 &= \frac{S_{XY}}{S_{YY}} = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(y_i-\overline{y})^2} \\
&\Rightarrow The\ slopes\ of\ two\ regression\ lines\ are\ different.
\end{aligned}
$$

**2.**

有可能是因為樣本數 $n$ 非常大而且 $variance$ 很大，造成每個變數的 $standard\ error$ 都非常小，使得在做檢測時的精準度非常高，因此才會在即使 $R^2$ 很小的情況下 (即模型對資料的解釋能力很低)，每個變數檢定時的 $p-value$ 依舊能達到非常顯著的程度。