20 Oct

6. Lack of fit H_0 : no lack of fit $\Rightarrow$ H_0 : model is fitted well

Let y_{ij} represent the jth response at the ith experimental combination, $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$

Example:

y_{ij} i = combination $i = 1, \ldots, m$; j = response for that combination $= 1, \ldots, n_i$

Table 3.1 Breadwrapper Stock Data

$n = 20$
$m = 15$
$p = 9$
$p' = 10$

y(g/in.)	x_1 (°F)	x_2 (°F)	x_3 (weight in %)		
6.6	225	46	0.5	$n_1 = 1$ y_{11}	$\hat{y}_1$
6.9	285	46	0.5	$n_2 = 1$ y_{21}	$\hat{y}_2$
7.9	225	64	0.5		
6.1	285	64	0.5		
9.2	225	46	1.7		
6.8	285	46	1.7		
10.4	225	64	1.7		
7.3	285	64	1.7		
9.8	204.5	55	1.1		
5.0	305.5	55	1.1		
6.9	255	39.9	1.1		
6.3	255	70.1	1.1		
4.0	255	55	0.09		
8.6	255	55	2.11	$n_{14} = 1$ $y_{14,1}$	$\hat{y}_{14}$
10.1	255	55	1.1		
9.9	255	55	1.1		
12.2	255	55	1.1		
9.7	255	55	1.1		
9.7	255	55	1.1		
9.6	255	55	1.1		

$\bar{y}_1 = y_{11}$

$\bar{y}_2 = y_{21}$

$\boxed{\text{repeated measurements on the same experimental combination}}$

We fit a model of x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, $x_2x_3 \Rightarrow p' = 10$.

$m = 15, n_1 = 1, n_2 = 1, \ldots, n_{14} = 1, n_{15} = 6$

$\sum_{j=1}^{m} n_i = n(= 20)$

$\boxed{m > p'}(= 10)$

$n_{15} = 6$ $\left.\begin{matrix} y_{15,1} \\ y_{15,2} \\ \vdots \\ y_{15,6} \end{matrix}\right\} \hat{y}_{15}$

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i + \bar{y}_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i - \hat{y}_i)^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(\bar{y}_i - \hat{y}_i)
\end{aligned}$$

$\bar{y}_{15} = \dfrac{y_{15,1} + \cdots + y_{15,6}}{6}$

Then

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

↑	↑	↑
Res S.S.	Pure error S.S.	Lack of fit S.S.

Res S.S. | fitted model

$$= (y_{11} - \hat{y}_1)^2 + (y_{21} - \hat{y}_2)^2 + \cdots + (y_{15,1} - \hat{y}_{15})^2 + (y_{15,2} - \hat{y}_{15})^2 + \cdots + (y_{15,6} - \hat{y}_{15})^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\underline{y_{ij} - \bar{y}_i} + \underline{\bar{y}_i - \hat{y}_i})^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2 + \underbrace{2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \hat{y}_i)}$$

$$= \cancel{2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)}$$

$$2 \sum_{i=1}^{m} (\bar{y}_i - \hat{y}_i) \boxed{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)} = 0$$

MATH3423

dist. of $\dfrac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{\sigma^2} \sim \chi^2_{(n_i - 1)}$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2$$

$$\Rightarrow \underbrace{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2}_{=\text{Res S.S.}\,|\,\text{fitted model}} = \underbrace{\sum_{i=1}^{m} \boxed{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}}_{=\text{Pure Error S.S.}} + \underbrace{\sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2}_{\text{Lack of fit S.S.}}$$

1 $\dfrac{\text{Res S.S.}\,|\,\text{fitted model}}{\sigma^2}$

$\dfrac{\text{Pure Error S.S.}}{\sigma^2} \sim \chi^2_{\boxed{\sum_{i=1}^{m}(n_i - 1)}}$

$\dfrac{\text{Lack of fit S.S.}}{\sigma^2}$

2 indep.

$$\sum_{i=1}^{m}(n_i - 1) = \sum_{i=1}^{m} n_i - m = n - m$$

n: # of ~~o~~total obs

m: # of distinct combinations of X

$\Rightarrow$ dist. of $\dfrac{\text{Lack of fit S.S.}}{\sigma^2}$

① $\dfrac{\text{Res S.S.} \mid \text{fitted model}}{\sigma^2}$

H_0 : no lack of fit

H_A : underfitted the model

unknown

True model : $\underset{\sim}{Y} = \underset{\sim}{X_1}\,\underset{\sim}{\beta_1} + \boxed{\underset{\sim}{X_2}\,\underset{\sim}{\beta_2}} + \underset{\sim}{e}$

underfitting

Fitted model : $\underset{\sim}{Y} = \underset{n\times p'}{\underset{\sim}{X_1}}\,\underset{\sim}{\beta_1} + \underset{\sim}{e}^{*}$ $n\times$

$\text{Res S.S.} \mid \text{fitted model} = \underset{\sim}{Y}^T(\underset{\sim}{I} - \underset{\sim}{X_1}(\underset{\sim}{X_1}^T\underset{\sim}{X_1})^{-1}\underset{\sim}{X_1}^T)\underset{\sim}{Y}$

By Theorem 3.4 in p.17 $\quad \underset{\sim}{Y} \sim MN(\underset{\sim}{\mu}, \underset{\sim}{I})$

$$\underset{\sim}{Y}^T\underset{\sim}{A}\,\underset{\sim}{Y} \sim \chi^2(\underset{\uparrow\ \text{rank of } \underset{\sim}{A}}{k}, \lambda)$$

non-centrality constant $\underset{\sim}{\mu}^T\underset{\sim}{A}\,\underset{\sim}{\mu}$

$$\frac{\text{Res S.S.} \mid \text{fitted model}}{\sigma^2} = \underset{\sim}{Y}^{*T}(\underset{\sim}{I} - \underset{\sim}{X_1}(\underset{\sim}{X_1}^T\underset{\sim}{X_1})^{-1}\underset{\sim}{X_1}^T)\underset{\sim}{Y}^{*}$$

where $\underset{\sim}{Y}^{*} = \underset{\sim}{Y}/\sigma \qquad \underset{\sim}{Y}^{*} \sim MN(\frac{\underset{\sim}{\mu}}{\sigma}, \underset{\sim}{I})$

$$\Uparrow$$

$$\underset{\sim}{\mu}^{*} = \frac{\underset{\sim}{\mu}}{\sigma} = \frac{\underset{\sim}{X_1}\,\underset{\sim}{\beta_1} + \underset{\sim}{X_2}\,\underset{\sim}{\beta_2}}{\sigma}$$

$k = \cancel{\text{p}}$ rank of $\underset{\sim}{A}$

$= \text{trace}\,(\underset{n\times n}{\underset{\sim}{I}} - \underbrace{\underset{n\times p'}{\underset{\sim}{X_1}}(\underset{\sim}{X_1}^T\underset{\sim}{X_1})^{-1}\underset{\sim}{X_1}^T}_{\uparrow})$

$\overset{\text{II}}{=} n - p'$

$\text{trace}\,(\underline{\underset{\sim}{X_1}(\underset{\sim}{X_1}^T\underset{\sim}{X_1})^{-1}}\,\underline{\underset{\sim}{X_1}^T})$

$= \cancel{\text{trace}\,(\underset{\sim}{I})}$

$= \text{trace}\,(\underline{\underset{\sim}{X_1}^T\underset{\sim}{X_1}}\ \underline{(\underset{\sim}{X_1}^T\underset{\sim}{X_1})^{-1}})$

$= \text{trace}\,(\underset{p'\times p'}{\underset{\sim}{I}})$

$$\lambda = \underset{\sim}{\mu}^{*T} \underset{\sim}{A}\, \underset{\sim}{\mu}^{*}$$

$$= \frac{1}{\sigma^2} (\underset{\sim}{X}_1 \beta_1 + \underset{\sim}{X}_2 \beta_2)^T (\underset{\sim}{I} - \underset{\sim}{X}_1 (\underset{\sim}{X}_1^T \underset{\sim}{X}_1)^{-1} \underset{\sim}{X}_1^T) (\underset{\sim}{X}_1 \beta_1 + \underset{\sim}{X}_2 \beta_2)$$

$$= \frac{1}{\sigma^2} \Big\{ ① \ \beta_1^T \underset{\sim}{X}_1^T (\underset{\sim}{I} - \underset{\sim}{X}_1 (\underset{\sim}{X}_1^T \underset{\sim}{X}_1)^{-1} \underset{\sim}{X}_1^T) \underset{\sim}{X}_1 \beta_1 \overset{=0}{} +$$

$$② \ \beta_2^T \underset{\sim}{X}_2^T (\underset{\sim}{I} - \underset{\sim}{X}_1 (\underset{\sim}{X}_1^T \underset{\sim}{X}_1)^{-1} \underset{\sim}{X}_1^T) \underset{\sim}{X}_1 \beta_1 \overset{=0}{} +$$

$$③ \ \beta_1^T \underset{\sim}{X}_1^T (\underset{\sim}{I} - \underset{\sim}{X}_1 (\underset{\sim}{X}_1^T \underset{\sim}{X}_1)^{-1} \underset{\sim}{X}_1^T) \underset{\sim}{X}_2 \beta_2 \overset{=0}{} +$$

$$④ \ \beta_2^T \underset{\sim}{X}_2^T (\underset{\sim}{I} - \underset{\sim}{X}_1 (\underset{\sim}{X}_1^T \underset{\sim}{X}_1)^{-1} \underset{\sim}{X}_1^T) \underset{\sim}{X}_2 \beta_2 \Big\}$$

$$= \frac{1}{\sigma^2} \beta_2^T (\underset{\sim}{X}_2^T \underset{\sim}{X}_2 - \underset{\sim}{X}_2^T \underset{\sim}{X}_1 (\underset{\sim}{X}_1^T \underset{\sim}{X}_1)^{-1} \underset{\sim}{X}_1^T \underset{\sim}{X}_2) \beta_2$$

$\neq 0$ except $\underset{\sim}{X}_2 = \underset{\sim}{0}$

$$\Rightarrow \frac{\text{Res S.S.} \mid \text{fitted model}}{\sigma^2} \sim \chi^2(n-p', \lambda)$$

$$E\left(\frac{\text{Res S.S.} \mid \text{fitted model}}{\sigma^2}\right) = n - p' + \lambda$$

$$\Rightarrow E\underbrace{\left(\frac{\text{Res S.S.} \mid \text{fitted model}}{n-p'}\right)}_{\boxed{\hat{\sigma}^2}\ \text{fitted model}} = \sigma^2 \left(1 + \frac{\lambda}{n-p'}\right) \neq \sigma^2$$

biased est. of σ^2 if the model is under-fitted.

residual mean squares

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 \qquad \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \qquad \sum_{i=1}^{m} n_i (\bar{y}_i - \boxed{\hat{y}_i})^2$$

$$\frac{\boxed{\text{Res S.S.} \mid \text{fitted model}}}{\sigma^2} = \frac{\boxed{\text{Pure Error S.S.}}}{\sigma^2} + \frac{\boxed{\text{Lack of fit S.S.}}}{\sigma^2}$$

$$\sim \chi^2(n-p', \lambda) \qquad \sim \chi^2_{\underbrace{\sum_{i=1}^{m}(n_i - 1)}_{n-m}}$$

not depend on the model — depend on the model

$$E\left(\frac{\text{Pure Error S.S.}}{\sigma^2}\right) = n-m$$

$$\Rightarrow E\left(\frac{\text{Pure Error S.S.}}{n-m}\right) = \sigma^2 \Rightarrow \frac{\text{Pure Error S.S.}}{n-m} \text{ is unbiased est. for } \sigma^2$$

② indep? Yes

$$\Rightarrow \frac{\text{Lack of fit S.S.}}{\sigma^2} \sim \chi^2(\underbrace{n-p'-\sum_{i=1}^{m}(n_i-1)}_{\substack{n-p'-\sum_{i=1}^{m} n_i + m \\ \shortparallel \\ m-p'}}, \lambda)$$

$$\lambda = \frac{1}{\sigma^2} \beta_2^T (X_2^T X_2 - X_2^T X_1 (X_1^T X_1)^{-1} X_1^T X_2) \beta_2$$

H_0 : No ~~#~~ lack of fit

$$\frac{\frac{\text{Lack of fit S.S.}}{\cancel{\sigma^2}} / (m-p')}{\frac{\text{Pure Error S.S.}}{\cancel{\sigma^2}} / (n-m)} \sim F(m-p', n-m, \lambda)$$

Under H_0,

$$\Rightarrow \frac{\text{Lack of fit S.S.} / (m-p')}{\text{Pure Error S.S.} (n-m)} \sim F(m-p', n-m)$$

$$\text{Test stat. } F = \frac{\sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2/(m-p')}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2/(n-m)}$$

Consider $E(F)$ under H_1

$$\Rightarrow \approx \frac{E\left(\frac{\sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2}{\sigma^2(m-p')}\right)}{E\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{\sigma^2(n-m)}\right)} = \frac{(m-p'+\lambda)/(m-p')}{(n-m)/(n-m)}$$

$$= 1 + \frac{\lambda}{m-p'}$$

where $\lambda = \frac{1}{\sigma^2} \beta_2^T X_2^T (I - X_1 (X_1^T X_1)^{-1} X_1^T) X_2 \beta_2$

$\lambda = 0$ under H_0

$\lambda > 0$ under $H_1 \Rightarrow E(F) > 1$

Reject H_0 if F_{obs} is significantly large

if $F_{obs} > F_\alpha(m-p', n-m)$

Calculation

① Res S.S. | fitted ~~model~~ model

② Calculate Pure Error SS. $= \sum_{i=1}^{m} \boxed{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}$

$= \sum_{i=1}^{m} (n_i - 1) \boxed{S_i^2}$

sample variance for the ith combination of $\underline{x}$

In our example, $S_1^2 = \cdots = S_{14}^2 = 0$

i.e. Find S_{15}^2 only

③ Lack of fit S.S. = Res S.S. | fitted model − Pure Error SS.

$$\mathrm{F} = \frac{\sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2/(m-p')}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-m)} \sim \mathrm{F}(m-p', n-m)$$

$$\text{Reject} H_0 \text{ if } \mathrm{F} > \mathrm{F}_\alpha(m-p', n-m)$$

$$\mathrm{E(F)} \approx \frac{\mathrm{E}(\sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2/(m-p'))}{\mathrm{E}(\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-m))}$$

$$= \frac{\sigma^2 + \frac{\beta_2^T(X_2^T X_2 - X_2^T X_1(X_1^T X_1)^{-1}X_1^T X_2)\beta_2}{m-p'}}{\sigma^2}$$

$$= 1 + \frac{\beta_2^T(X_2^T X_2 - X_2^T X_1(X_1^T X_1)^{-1}X_1^T X_2)\beta_2}{\sigma^2(m-p')}$$

sample variance based on all y_{ij}

Fit a model of y on $x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3$

Reg S.S. = Total S.S. − Residual S.S. Total S.S. $= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$

$= (n-1)S^2$ OR $= S_{yy}$

Table 3.1

H_0 = no lack-of-fit

Source	S.S.	d.f.	M.S.	F
Regression	70.302	9	7.8113	⊛
(given) Residual → Error	11.8678	10	1.18678	
Lack of fit	6.9078	5	1.3816	1.39 $= \frac{6.9078/5}{4.96/5}$
Pure Error	4.96	5	0.9920	
Total	82.17	19 ← $n-1$		

Pure Error S.S. $= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$

$= \sum_{i=1}^{m}(n_i - 1)S_i^2$

$= (6-1)S_{15}^2$

10.1, 9.9, 12.2, ..., 9.6

$= 4.96$

d.f. for Pure Error $= \sum_{i=1}^{m}(n_i - 1) = 5$

Lack of fit S.S. = Residual S.S. | fitted model − Pure Error S.S.

$= 11.8678 - 4.96$

$= 6.9078$

d.f. for Lack of fit S.S. = d.f. for Res S.S. | fitted model − d.f. for Pure Error S.S.

$= 10 - 5 = 5$

$m - p' = 15 - 10$

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_{1sq} = \beta_{2sq} = \beta_{3sq} = \beta_{12} = \beta_{13} = \beta_{23} = 0$

⊛ $\mathrm{F} = \frac{\text{Reg M.S.}}{\text{Residual M.S.}} =$

- Can't reject H_0: ~~to~~ no lack of fit

$$\Rightarrow \hat{\sigma}^2 = \frac{\text{Res S.S.} \mid \text{fitted model}}{n-p'} \quad \text{— unbiased est. of } \sigma^2$$

- However, Reject H_0

$$\hat{\sigma}^2 = \frac{\text{Res S.S.} \mid \text{fitted model}}{n-p'} \quad \text{— biased est. of } \sigma^2$$

$$\Rightarrow \hat{\sigma}^2_{\text{pure error}} = \frac{\text{Pure Error S.S.}}{n-m}$$

— unbiased est. of σ^2

Based on $\hat{\sigma}^2 \mid$ pure error to perform hypothesis testing & construct the confidence interval.

Q: Is $\hat{\beta}$ unbiased?

e.g. Fitted $y_i = \beta_0 + \beta_1 X_{i1} + e_i$ $\Rightarrow \hat{\beta}_0 \mid \text{fitted}, \hat{\beta}_1 \mid \text{fitted}$

True $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$

$$\hat{\beta}_1 \mid \text{fitted} = \frac{S_{x_1 y}}{S_{x_1 x_1}}$$

$$E(\hat{\beta}_1 \mid \text{fitted}) = E\left(\frac{S_{x_1 y}}{S_{x_1 x_1}}\right)$$

$$= \frac{1}{S_{x_1 x_1}} E\left(\sum_{i=1}^n (X_{i1} - \bar{X}_1) y_i\right)$$

$$= \frac{1}{S_{x_1 x_1}} \cancel{E(}\sum_{i=1}^n (X_{i1} - \bar{X}_1) \underbrace{\boxed{E(y_i)}}_{= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}}$$

correlation coeff. of X_1 & $X_2 = 0$

$$= \frac{1}{S_{x_1 x_1}} \Big[\beta_0 \underbrace{\boxed{\sum_{i=1}^n (X_{i1} - \bar{X}_1)}}_{=0} + \beta_1 \sum_{i=1}^n (X_{i1} - \bar{X}_1) X_{i1} + \underbrace{\beta_2 \sum_{i=1}^n (X_{i1} - \bar{X}_1) X_{i2}}_{\text{except } =0}\Big]$$

$$= \frac{1}{S_{x_1 x_1}} [\beta_1 S_{x_1 x_1} + \beta_2 \boxed{S_{x_1 x_2}}]$$

— biased est.