# ANOVA
# Analysis of Variances

Argon Chen
Industrial Engineering
National Taiwan University



## Replicated $2^2$ Factorial Design and Sum of Squares (SS's)

| Test | $x_1$ | $x_2$ | $y_{i1}$ | $y_{i2}$ | $y_{i3}$ | $y_{i4}$ | $y_{i5}$ | $\bar{y}_i$ |
|---|---|---|---|---|---|---|---|---|
| 1 | −1 | −1 | 11 | 7 | 10 | 15 | 7 | 10 |
| 2 | +1 | −1 | 48 | 43 | 52 | 55 | 47 | 49 |
| 3 | −1 | +1 | 31 | 24 | 27 | 23 | 20 | 25 |
| 4 | +1 | +1 | 37 | 33 | 34 | 37 | 34 | 35 |

$i = 1,\ldots,m \quad m = \#\text{ of Tests} = 4 \qquad j = 1,\ldots,n \quad n = \#\text{ of replicates} = 5$

$$\bar{y}_i = \frac{\sum_{j=1}^{n} y_{ij}}{n} \quad i = 1,\ldots,m \qquad \bar{\bar{y}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}}{mn} = 29.75$$

$$y_{ij} = \bar{\bar{y}} + (\bar{y}_i - \bar{\bar{y}}) + (y_{ij} - \bar{y}_i) \qquad \text{Sum of Squares=SS} = \sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}^2$$

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

$$\text{SS} = \text{SS(mean)} + \text{SS(between tests)} + \text{SS(within tests)}$$

# Sum of Squares

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

$$\text{SS} - \text{SS(mean)} = \text{SS(total)} = \text{SS(between tests)} + \text{SS(within tests)}$$

$$\text{SS(total)} = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}^2 - mn\bar{\bar{y}}^2$$

$$= 21969 - 17701.25$$

$$\text{SS(between tests)} = n\sum_{i=1}^{m}(\bar{y}_i-\bar{\bar{y}})^2 = 4053.75$$

$$\text{SS(within tests)} = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2 = 214$$

$$= \text{Sum of Squared Errors} = \text{SSE}$$



# Mean Squares

- Mean squares within tests: Mean Squared Error (MSE)

$$\hat{\sigma}_\varepsilon^2 = s_p^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2}{m\quad(n-1)} = \frac{\text{SS(within tests)}}{m(n-1)} = \frac{\text{SS(within tests)}}{\text{degrees of freedom}}$$

$$= \frac{214}{4(5-1)} = 13.375 = \text{Mean Square (within tests)=MS(within tests)}$$

$$= \text{Mean Squared Error} = \text{MSE}$$

- Mean squares between tests: MS(between tests)

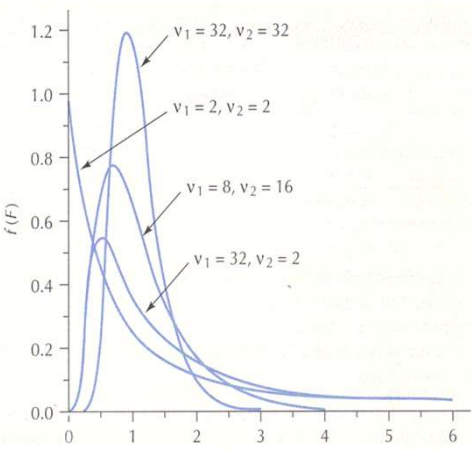
$$s_{\bar{y}}^2 = \frac{\sum_{i=1}^{m}(\bar{y}_i-\bar{\bar{y}})^2}{m-1} \Rightarrow s_{\bar{y}}^2 = \frac{s_y^2}{n} \Rightarrow s_y^2 = ns_{\bar{y}}^2$$ $n$ is the sample size of $\bar{y}_i$

$$s_y^2 = ns_{\bar{y}}^2 = \frac{n\sum_{i=1}^{m}(\bar{y}_i-\bar{\bar{y}})^2}{m-1} = \frac{\text{SS(between tests)}}{\text{DOF(between tests)}}$$

$$= \frac{4053.75}{4-1} = 1351.25 = \text{MS(between tests)}$$

# Recall *F* Statistic and Distributions

- Two ***normal*** populations have the ***same variance***
- $s_1^2$ : sample variances from population 1 with sample size $n_1$
- $s_2^2$ : sample variances from population 2 with sample size $n_2$
- Then, $s_1^2 / s_2^2$ follows $F_{v1, v2}$ distribution
  - $v_1=n_1-1$; $v_2=n_2-1$
- F table look-up




5

# F-test

- $H_0$: *MS*(between tests)=*MS*(within tests)
  $H_a$: *MS*(between tests)>*MS*(within tests)
- Test statistic:
  $F_{calc}$=*MS*(between tests)/*MS*(within tests)
- Reject $H_0$ when $F_{calc}>F_{v1,v2,1-\alpha}$
- Reject $H_0$: **Difference between tests are not caused by noise (estimated by difference within tests)**
- Example:

$$F_{calc} = \frac{MS(\text{between tests})}{MS(\text{within tests})} = \frac{1351.25}{13.375} = 101.03 > F_{3,16,0.99} = 5.29$$

6

# ANOVA Table for Replicated $2^2$

| Source of variation | Sum of Squares (*SS*) | Degrees of Freedom *(DOF)* | Mean square (MS) | *F Ratio* |
|---|---|---|---|---|
| Mean | $mn\bar{\bar{y}}^2$ | 1 | SS(Mean)/ DoF(Mean) | MS(Mean)/MS(P ure error) |
| Between tests | $n\sum_{i=1}^{m}(\bar{y}_i - \bar{\bar{y}})^2$ | *m*–1 | SS(B/W Tests)/DoF(B/W Tests) | MS(B/w Tests)/MS(Pure error) |
| Pure error (within tests) | $\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$ | *m*(*n*–1) | SS(Pure error)/DoF(Pure error) | |
| SS | $\sum_{i=1}^{m}\sum_{j=1}^{n}y_{ij}^2$ | *mn* | | |

| Source of variation | Sum of Squares | *DOF* | Mean square | $F_{calc}$ |
|---|---|---|---|---|
| Mean | 17,701.25 | 1 | 17,701.25 | 1,323.46 |
| Between tests | 4,053.75 | 3 | 1,351.25 | 101.03 |
| Pure error (within tests) | 214.00 | 16 | 13.375 | |
| SS | **21,969.00** | **20** | | |

7

# ANOVA Table without "Mean"

$$\mathrm{SS(total)} = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}y_{ij}^2 - mn\bar{\bar{y}}^2$$

| Source of variation | Sum of Squares (*SS*) | Degrees of Freedom *(DOF)* | Mean square (MS) | *F Ratio* |
|---|---|---|---|---|
| Between tests | $n\sum_{i=1}^{m}(\bar{y}_i - \bar{\bar{y}})^2$ | *m*–1 | SS(B/W Tests)/DoF(B/W Tests) | MS(B/w Tests)/MS(Pure error) |
| Pure error (within tests) | $\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$ | *m*(*n*–1) | SS(Pure error)/DoF(Pure error) | |
| SS(Total)= SS-mean | $\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{\bar{y}})^2$ | *mn*–1 | | |

8

# ANOVA Table without "Mean"

$$SS(\text{total}) = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}^2 - mn\bar{\bar{y}}^2$$

$$= 21969 - 17701.25 = 4267.75$$

| Source of variation | Sum of Squares | *DOF* | Mean square | $F_{calc}$ |
|---|---|---|---|---|
| Between tests | 4,053.75 | 3 | 1,351.25 | 101.03 |
| Pure error (within tests) | 214.00 | 16 | 13.375 | |
| SS-Mean =SS(Total) | **4,267.75** | **19** | | |

9

# Linear Regression Model

$$\hat{y} = 29.75 + 12.25x_1 + 0.25x_2 - 7.25x_1x_2$$

| | X1 | X2 | X1X2 | Y |
|---|---|---|---|---|
| Replicate 1 | -1 | -1 | 1 | 11 |
| | 1 | -1 | -1 | 48 |
| | -1 | 1 | -1 | 31 |
| | 1 | 1 | 1 | 37 |
| Replicate 2 | -1 | -1 | 1 | 7 |
| | 1 | -1 | -1 | 43 |
| | -1 | 1 | -1 | 24 |
| | 1 | 1 | 1 | 33 |
| Replicate 3 | -1 | -1 | 1 | 10 |
| | 1 | -1 | -1 | 52 |
| | -1 | 1 | -1 | 27 |
| | 1 | 1 | 1 | 34 |
| Replicate 4 | -1 | -1 | 1 | 15 |
| | 1 | -1 | -1 | 55 |
| | -1 | 1 | -1 | 23 |
| | 1 | 1 | 1 | 37 |
| Replicate 5 | -1 | -1 | 1 | 7 |
| | 1 | -1 | -1 | 47 |
| | -1 | 1 | -1 | 20 |
| | 1 | 1 | 1 | 34 |

| 迴歸統計 | |
|---|---|
| R 的倍數 | 0.974606 |
| R 平方 | 0.949856 |
| 調整的 R | 0.940455 |
| 標準誤 | 3.657185 |
| 觀察值個 | 20 |

ANOVA

| | 自由度 | SS | MS | F | 顯著值 |
|---|---|---|---|---|---|
| 迴歸 | 3 | 4053.75 | 1351.25 | 101.028 | 1.3E-10 |
| 殘差 | 16 | 214 | 13.375 | | |
| 總和 | 19 | 4267.75 | | | |

| | 係數 | 標準誤 | t 統計 | P-值 | 下限 95% | 上限 95% |
|---|---|---|---|---|---|---|
| 截距 | 29.75 | 0.817771 | 36.37936 | 8.18E-17 | 28.0164 | 31.4836 |
| X1 | 12.25 | 0.817771 | 14.97974 | 7.8E-11 | 10.5164 | 13.9836 |
| X2 | 0.25 | 0.817771 | 0.305709 | 0.763768 | -1.4836 | 1.983597 |
| X1X2 | -7.25 | 0.817771 | -8.86556 | 1.43E-07 | -8.9836 | -5.5164 |

10

# $R^2$ to Assess the Model

Toal Variation in y SST= SSR + SSE

Overall variability in y

Explained in part by → The regression model

Remains, in part, unexplained → The error

- $R^2$ measures the proportion of the variation in y that is explained by the variation in x.

$$R^2 = \frac{\text{Variation explained by Model}}{\text{Total variation in y}} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

11

# Example of $R^2$

$$R = \sqrt{R^2} = \sqrt{0.949856} = 0.974606$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{4053.75}{4267.75} = 0.949856$$

| 迴歸統計 | |
|---|---|
| R 的倍數 | 0.974606 |
| R 平方 | 0.949856 |
| 調整的 R | 0.940455 |
| 標準誤 | 3.657185 |
| 觀察值個 | 20 |

ANOVA

| | 自由度 | SS | MS | F | 顯著值 |
|---|---|---|---|---|---|
| 迴歸 | 3 | 4053.75 | 1351.25 | 101.028 | 1.3E-10 |
| 殘差 | 16 | 214 | 13.375 | | |
| 總和 | 19 | 4267.75 | | | |

| | 係數 | 標準誤 | t 統計 | P-值 | 下限 95% | 上限 95% |
|---|---|---|---|---|---|---|
| 截距 | 29.75 | 0.817771 | 36.37936 | 8.18E-17 | 28.0164 | 31.4836 |
| X1 | 12.25 | 0.817771 | 14.97974 | 7.8E-11 | 10.5164 | 13.9836 |
| X2 | 0.25 | 0.817771 | 0.305709 | 0.763768 | -1.4836 | 1.983597 |
| X1X2 | -7.25 | 0.817771 | -8.86556 | 1.43E-07 | -8.9836 | -5.5164 |

12

# Adjusted $R^2$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \Rightarrow \text{Adjusted } R^2 = 1 - \frac{SSE/DoF}{SST/DoF}$$

| 迴歸統計 | |
|---|---|
| R 的倍數 | 0.974606 |
| R 平方 | 0.949856 |
| 調整的 R | 0.940455 |
| 標準誤 | 3.657185 |
| 觀察值個 | 20 |

$$\text{Adjusted } R^2 = 1 - \frac{214/16}{4267.75/19} = 0.940455$$

ANOVA

| | 自由度 | SS | MS | F | 顯著值 |
|---|---|---|---|---|---|
| 迴歸 | 3 | 4053.75 | 1351.25 | 101.028 | 1.3E-10 |
| 殘差 | 16 | 214 | 13.375 | | |
| 總和 | 19 | 4267.75 | | | |

| | 係數 | 標準誤 | t 統計 | P-值 | 下限 95% | 上限 95% |
|---|---|---|---|---|---|---|
| 截距 | 29.75 | 0.817771 | 36.37936 | 8.18E-17 | 28.0164 | 31.4836 |
| X1 | 12.25 | 0.817771 | 14.97974 | 7.8E-11 | 10.5164 | 13.9836 |
| X2 | 0.25 | 0.817771 | 0.305709 | 0.763768 | -1.4836 | 1.983597 |
| X1X2 | -7.25 | 0.817771 | -8.86556 | 1.43E-07 | -8.9836 | -5.5164 |

13

# Two-Way ANOVA

- Example: agriculture experiments
  - Three tomato varieties (*i*): Harvester, Pusa Early Dwarf and Ife No. 1
  - Four plantation densities (*j*): 10, 20, 30, and 40 thousand plants per hectare)
  - Three replicates (*k*)
- Experimental results: $x_{ijk}$

| | Plantation Density | | | | | |
|---|---|---|---|---|---|---|
| Variety | 10k | 20k | 30k | 40k | sum | $X_{i\bullet\bullet}$ |
| H | 10.5, 9.2, 7.9 | 12.8, 11.2, 13.3 | 12.1, 12.6, 14.0 | 10.8, 9.1, 12.5 | 136.0 | 11.3 |
| Ife | 8.1, 8.6, 10.1 | 12.7, 13.7, 11.5 | 14.4, 15.4, 13.7 | 11.3, 12.5, 14.5 | 146.5 | 12.21 |
| P | 16.1, 15.3, 17.5 | 16.6, 19.2, 18.5 | 20.8, 18.0, 21.0 | 18.4, 18.9, 17.2 | 217.5 | 18.13 |
| sum | 103.3 | 129.5 | 142.0 | 125.2 | 500.00 | |
| $X_{\bullet j\bullet}$ | 11.48 | 14.39 | 15.78 | 13.91 | | 13.89 |

14

# Sum of Squares

$i = 1, \cdots, I$

$j = 1, \cdots, J$

$k = 1, \cdots, K$

$N$= # of total trials(=$IJK$)

$X_{\bullet\bullet\bullet} = \sum_i \sum_j \sum_k X_{ijk} / N$

$X_{ij\bullet} = \sum_k X_{ijk} / K$

$X_{i\bullet\bullet} = \sum_j \sum_k X_{ijk} / (N / I)(=JK)$

$X_{\bullet j\bullet} = \sum_i \sum_k X_{ijk} / (N / J)(=IK)$

$$\text{SS(Total)} = \sum_i \sum_j \sum_k (X_{ijk} - X_{\bullet\bullet\bullet})^2$$

$$\text{SSE} = \sum_i \sum_j \sum_k (X_{ijk} - X_{ij\bullet})^2$$

$$\text{SS(Variety)} = \sum_i \sum_j \sum_k (X_{i\bullet\bullet} - X_{\bullet\bullet\bullet})^2 = \frac{N}{I} \sum_i (X_{i\bullet\bullet} - X_{\bullet\bullet\bullet})^2$$

$$\text{SS(Density)} = \sum_i \sum_j \sum_k (X_{\bullet j\bullet} - X_{\bullet\bullet\bullet})^2 = \frac{N}{J} \sum_j (X_{\bullet j\bullet} - X_{\bullet\bullet\bullet})^2$$

$$\text{SS(Interaction)} = \sum_i \sum_j \sum_k [(X_{ij\bullet} - X_{\bullet\bullet\bullet}) - (X_{i\bullet\bullet} - X_{\bullet\bullet\bullet}) - (X_{\bullet j\bullet} - X_{\bullet\bullet\bullet})]^2$$

$$= \sum_i \sum_j \sum_k (X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet})^2 = \frac{N}{IJ} \sum_i \sum_j (X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet})^2$$

$$\text{SS(Total)} = \text{SS(Variety)} + \text{SS(Density)} + \text{SS(Interaction)} + \text{SSE}$$

15

# Mean Squares

$$\text{MSE} = \sum_i \sum_j \sum_k (X_{ijk} - X_{ij\bullet})^2 / (K-1)IJ$$

$$\text{MS(Variety)} = \sum_i \sum_j \sum_k (X_{i\bullet\bullet} - X_{\bullet\bullet\bullet})^2 / (I-1) = \frac{N}{I} \sum_i (X_{i\bullet\bullet} - X_{\bullet\bullet\bullet})^2 / (I-1)$$

$$\text{MS(Density)} = \sum_i \sum_j \sum_k (X_{\bullet j\bullet} - X_{\bullet\bullet\bullet})^2 / (J-1) = \frac{N}{J} \sum_j (X_{\bullet j\bullet} - X_{\bullet\bullet\bullet})^2 / (J-1)$$

$$\text{MS(Interaction)} = \sum_i \sum_j \sum_k (X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet})^2 / (I-1)(J-1)$$

$$= \frac{N}{IJ} \sum_i \sum_j (X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet})^2 / (I-1)(J-1)$$

16

# ANOVA Table

| Source of variation | Sum of Squares | *DOF* | Mean square | $F_{calc}$ |
|---|---|---|---|---|
| Variety | 327.60 | 2 | 163.8 | 163.8/1.59 |
| Density | 86.69 | 3 | 28.9 | 28.9/1.59 |
| Interaction | 8.03 | 6 | 1.34 | 1.34/1.59 |
| Pure error (within tests) | 38.04 | 24 | 1.59 | |
| Total | 460.36 | 35 | | |

$$F_{2,24,0.99} = 5.61 \;\; F_{3,24,0.99} = 4.241 \;\; F_{6,24,0.99} = 3.63$$

17

# Two-Way ANOVA in Excel

| | | Plantation Density | | | |
|---|---|---|---|---|---|
| | | 10K | 20K | 30K | 40K |
| Variety | Harvester | 10.5 | 12.8 | 12.1 | 10.8 |
| | | 9.2 | 11.2 | 12.6 | 9.1 |
| | | 7.9 | 13.3 | 14 | 12.5 |
| | Ife No.1 | 8.1 | 12.7 | 14.4 | 11.3 |
| | | 8.6 | 13.7 | 15.4 | 12.5 |
| | | 10.1 | 11.5 | 13.7 | 14.5 |
| | Pusa Early Dwarf | 16.1 | 16.6 | 20.8 | 18.4 |
| | | 15.3 | 19.2 | 18 | 18.9 |
| | | 17.5 | 18.5 | 21 | 17.2 |

| ANOVA | | | | | | |
|---|---|---|---|---|---|---|
| 變源 | SS | 自由度 | MS | F | P-值 | 臨界值 |
| 樣本 | 327.5972 | 2 | 163.7986 | 103.343 | 1.61E-12 | 3.402826 |
| 欄 | 86.68667 | 3 | 28.89556 | 18.23063 | 2.21E-06 | 3.008787 |
| 交互作用 | 8.031667 | 6 | 1.338611 | 0.84455 | 0.548361 | 2.508189 |
| 組內 | 38.04 | 24 | 1.585 | | | |
| | | | | | | |
| 總和 | 460.3556 | 35 | | | | |

18

# Additive Model with Interactions

Let $\mu = \sum_i \sum_j \mu_{ij} / IJ \quad \mu_{i\bullet} = \sum_j \mu_{ij} / J \quad \mu_{\bullet j} = \sum_i \mu_{ij} / I$

$\alpha_i = \mu_{i\bullet} - \mu$ = effect of factor *A* at level *i*

$\beta_j = \mu_{\bullet j} - \mu$ = effect of factor *B* at level *j*

$\gamma_{ij} = (\mu_{ij} - \mu) - \alpha_i - \beta_j$ = interaction effect of *A* at level *i* and *B* at level *j*

Then, the additive linear model is:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

19

# ANOVA Table for All Terms (Back to Text-book Example)

| Source of variation | Sum of Squares | *DOF* | Mean square | $F_{calc}$ |
|---|---|---|---|---|
| $E_1$ | 3,001.25 | 1 | 3,001.25 | **224.39*** |
| $E_2$ | 1.25 | 1 | 1.25 | 0.09 |
| $E_{12}$ | 1,051.25 | 1 | 1,051.25 | **78.60*** |
| Pure error (within tests) | 214.00 | 16 | 13.375 | |
| Total | **4,267.75** | **19** | | |

20

# ANOVA Table for Partial Model

- Partial model: $\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_{12} x_1 x_2$

| Source of variation | Sum of Squares | *DOF* | Mean square | $F_{calc}$ |
|---|---|---|---|---|
| $E_1$ | 3,001.25 | 1 | 3,001.25 | **224.39*** |
| $E_{12}$ | 1,051.25 | 1 | 1,051.25 | **78.60*** |
| Residual $E_2$ | 1.25 | 1 | 1.25 | 0.09 |
| Pure error | 214.00 | 16 | 13.375 | |
| Total | **4,267.75** | **19** | | |

$$R^2 = \frac{SSR}{SST} = \frac{4052.5}{4267.75} = 0.949564$$



# Linear Regression for the Partial Model

| | X1 | X1X2 | Y |
|---|---|---|---|
| Replicate | -1 | 1 | 11 |
| | 1 | -1 | 48 |
| | -1 | -1 | 31 |
| | 1 | 1 | 37 |
| Replicate | -1 | 1 | 7 |
| | 1 | -1 | 43 |
| | -1 | -1 | 24 |
| | 1 | 1 | 33 |
| Replicate | -1 | 1 | 10 |
| | 1 | -1 | 52 |
| | -1 | -1 | 27 |
| | 1 | 1 | 34 |
| Replicate | -1 | 1 | 15 |
| | 1 | -1 | 55 |
| | -1 | -1 | 23 |
| | 1 | 1 | 37 |
| Replicate | -1 | 1 | 7 |
| | 1 | -1 | 47 |
| | -1 | -1 | 20 |
| | 1 | 1 | 34 |

| 迴歸統計 | |
|---|---|
| R 的倍數 | 0.974456 |
| R 平方 | 0.949564 |
| 調整的 R | 0.94363 |
| 標準誤 | 3.558337 |
| 觀察值個 | 20 |

$$R^2 = \frac{SSR}{SST} = \frac{4052.5}{4267.75} = 0.949564$$

ANOVA

| | 自由度 | SS | MS | F | 顯著值 |
|---|---|---|---|---|---|
| 迴歸 | 2 | 4052.5 | 2026.25 | 160.029 | 9.4E-12 |
| 殘差 | 17 | 215.25 | 12.66176 | | |
| 總和 | 19 | 4267.75 | | | |

| | 係數 | 標準誤 | t 統計 | P-值 | 下限 95% | 上限 95% | 下限 95.0% |
|---|---|---|---|---|---|---|---|
| 截距 | 29.75 | 0.795668 | 37.38995 | 9.12E-18 | 28.07128 | 31.42872 | 28.07128 |
| X1 | 12.25 | 0.795668 | 15.39586 | 2.05E-11 | 10.57128 | 13.92872 | 10.57128 |
| X1X2 | -7.25 | 0.795668 | -9.11184 | 5.94E-08 | -8.92872 | -5.57128 | -8.92872 |

# Sum of Squares (Interactions)

$$\text{SS}(ij \text{ interaction}) = \frac{N}{IJ}\sum_i\sum_j [X_{ij\bullet\bullet} - X_{i\bullet\bullet\bullet} - X_{\bullet j\bullet\bullet} + X_{\bullet\bullet\bullet\bullet}]^2$$

$$= \frac{N}{IJ}\sum_i\sum_j [(X_{ij\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{i\bullet\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet j\bullet\bullet} - X_{\bullet\bullet\bullet\bullet})]^2$$

$$\text{SS}(il \text{ interaction}) = \frac{N}{IL}\sum_i\sum_l [X_{i\bullet l\bullet} - X_{i\bullet\bullet\bullet} - X_{\bullet\bullet l\bullet} + X_{\bullet\bullet\bullet\bullet}]^2$$

$$= \frac{N}{IL}\sum_i\sum_l [(X_{i\bullet l\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{i\bullet\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet\bullet l\bullet} - X_{\bullet\bullet\bullet\bullet})]^2$$

$$\text{SS}(jl \text{ interaction}) = \frac{N}{JL}\sum_j\sum_l [X_{\bullet jl\bullet} - X_{\bullet j\bullet\bullet} - X_{\bullet\bullet l\bullet} + X_{\bullet\bullet\bullet\bullet}]$$

$$= \frac{N}{JL}\sum_j\sum_l [(X_{\bullet jl\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet j\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet\bullet l\bullet} - X_{\bullet\bullet\bullet\bullet})]^2$$

$\text{SS}(ijl \text{ interaction})$

$$= K\sum_i\sum_j\sum_l \{(X_{ijl\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{i\bullet\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet j\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet\bullet l\bullet} - X_{\bullet\bullet\bullet\bullet})$$
$$- [(X_{ij\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{i\bullet\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet j\bullet\bullet} - X_{\bullet\bullet\bullet\bullet})]$$
$$- [(X_{i\bullet l\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{i\bullet\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet\bullet l\bullet} - X_{\bullet\bullet\bullet\bullet})]$$
$$- [(X_{\bullet jl\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet j\bullet\bullet} - X_{\bullet\bullet\bullet\bullet}) - (X_{\bullet\bullet l\bullet} - X_{\bullet\bullet\bullet\bullet})]\}^2$$

$$= K\sum_i\sum_j\sum_l [X_{ijl\bullet} - X_{ij\bullet\bullet} - X_{i\bullet l\bullet} - X_{\bullet jl\bullet} + X_{i\bullet\bullet\bullet} + X_{\bullet j\bullet\bullet} + X_{\bullet\bullet l\bullet} - X_{\bullet\bullet\bullet\bullet}]^2$$

$i = \mathbf{1}, \cdots, I$ (Factor 1)

$j = \mathbf{1}, \cdots, J$ (Factor 2)

$l = \mathbf{1}, \cdots, L$ (Factor 3)

$k = \mathbf{1}, \cdots, K$ (Replicates)

$N$=# *of total trials*(=$IJLK$)

$$X_{\bullet\bullet\bullet\bullet} = \sum_i\sum_j\sum_l\sum_k X_{ijlk} / N$$

$$X_{i\bullet\bullet\bullet} = \sum_j\sum_l\sum_k X_{ijlk} / (N/I)(= JLK)$$

$$X_{ij\bullet\bullet} = \sum_l\sum_k X_{ijlk} / (N/IJ)(= LK)$$

$$X_{ijl\bullet} = \sum_k X_{ijlk} / K$$



# Analysis of Surface Defects Data

| Factor | Average η by Factor Level (dB) 1 | 2 | 3 | Degree of Freedom | Sum of Squares | Mean Square | F |
|---|---|---|---|---|---|---|---|
| A. Temperature | − 24.23 | −50.10 | − 61.76 | 2 | 4427 | 2214 | 27 |
| B. Pressure | − 27.55 | −47.44 | − 61.10 | 2 | 3416 | 1708 | 21 |
| C. Nitrogen | −39.03 | − 55.99 | − 41.07 | 2 | 1030 | 515 | 6.4 |
| D. Silane | − 39.20 | − 46.85 | −50.04 | 2 | 372 | 186 | 2.3 |
| E. Settling time | −51.52 | − 40.54 | − 44.03 | 2 | 378 | 189 | 2.3 |
| F. Cleaning method | −45.56 | − 41.58 | − 48.95 | 2 | 164† | 82 | |
| Error | | | | 5 | 405† | 81 | |
| Total | | | | 17 | 10192 | | |
| (Error) | | | | (7) | (569) | (81) | |

* Overall mean η = −45.36 dB. Underscore indicates starting level.

† Indicates the sum of squares added together to form the pooled error sum of squares shown in parentheses.

# Analysis of Thickness Data

| Factor | Average η′ by Level (dB) | | | Degree of Freedom | Sum of Squares | Mean Square | F |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | | | | |
| A. Temperature | 35.12 | 34.91 | 24.52 | 2 | 440 | 220 | 16 |
| B. Pressure | 31.61 | 30.70 | 32.24 | 2 | 7† | 3.5 | |
| C. Nitrogen | 34.39 | 27.86 | 32.30 | 2 | 134 | 67 | 5.0 |
| D. Silane | 31.68 | 34.70 | 28.17 | 2 | 128 | 64 | 4.8 |
| E. Settling time | 30.52 | 32.87 | 31.16 | 2 | 18† | 9 | |
| F. Cleaning method | 27.04 | 33.67 | 33.85 | 2 | 181 | 90.5 | 6.8 |
| Error | | | | 5 | 96† | 19.2 | |
| Total | | | | 17 | 1004 | 59.1 | |
| (Error) | | | | (9) | (121) | (13.4) | |

* Overall mean η′ = 31.52 dB. Underscore indicates starting level.
† Indicates the sum of squares added together to form the pooled error sum of squares shown in parentheses.

25

# Analysis of Deposition Rate Data

| Factor | Average η″ by Factor Level (dBam) | | | Degree of Freedom | Sum of Squares | Mean Square | F |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | | | | |
| A. Temperature | 28.76 | 34.13 | 39.46 | 2 | 343.1 | 171.5 | 553 |
| B. Pressure | 32.03 | 34.78 | 35.54 | 2 | 41.0 | 20.5 | 66 |
| C. Nitrogen | 32.81 | 35.29 | 34.25 | 2 | 18.7 | 9.4 | 30 |
| D. Silane | 32.21 | 34.53 | 35.61 | 2 | 36.3 | 18.1 | 58 |
| E. Settling time | 34.06 | 33.99 | 34.30 | 2 | 0.3† | 0.2 | |
| F. Cleaning method | 33.81 | 34.10 | 34.44 | 2 | 1.2† | 0.6 | |
| Error | | | | 5 | 1.3† | 0.26 | |
| Total | | | | 17 | 441.9 | 25.9 | |
| (Error) | | | | (9) | (2.8) | (0.31) | |

* Overall mean η″ = 34.12 dBam. Underscore indicates starting level.
† Indicates the sum of squares added together to form the pooled error sum of squares shown in parentheses.

26

# More on Linear Regression

Argon Chen
Industrial Engineering
National Taiwan University



# Multiple Regression?

**Example: Fuel Consumption**

Y=FUEL=1000×FUELC/POP
= motor fuel consumption, gallons per person

$X_1$=TAX, cents per gallon

$X_2$=DLIC=NLIC/POP
=proportion of population with driver's licenses.

Six columns of Table 2.1 list values, for each of the 48 contiguous states, of the following quantities:

POP = 1971 population (in thousands).
TAX = 1972 motor fuel tax in cents per gallon.
NLIC = 1971 number of licensed drivers (in thousands).
INC = 1972 per capita personal income in dollars.
ROAD = 1971 length of Federal-aid primary highways, in miles.

| STATE | POP | $X_1$ TAX | NLIC | $X_3$ INC | $X_4$ ROAD | FUELC | $X_2$ DLIC | $Y$ FUEL |
|---|---|---|---|---|---|---|---|---|
| 1 ME | 1029 | 9.00 | 540 | 3571 | 1976 | 557 | 0.525 | 541 |
| 2 NH | 771 | 9.00 | 441 | 4092 | 1250 | 404 | 0.572 | 524 |
| 3 VT | 462 | 9.00 | 268 | 3865 | 1586 | 259 | 0.580 | 561 |
| 4 MA | 5787 | 7.50 | 3060 | 4870 | 2351 | 2396 | 0.529 | 414 |
| 5 RI | 968 | 8.00 | 527 | 4399 | 431 | 397 | 0.544 | 410 |
| 6 CN | 3082 | 10.00 | 1760 | 5342 | 1333 | 1408 | 0.571 | 457 |
| 7 NY | 18366 | 8.00 | 8278 | 5319 | 11868 | 6312 | 0.451 | 344 |
| 8 NJ | 7367 | 8.00 | 4074 | 5126 | 2138 | 3439 | 0.553 | 467 |
| 9 PA | 11926 | 8.00 | 6312 | 4447 | 8577 | 5528 | 0.529 | 464 |
| 10 OH | 10783 | 7.00 | 5948 | 4512 | 8507 | 5375 | 0.552 | 498 |
| 11 IN | 5291 | 8.00 | 2804 | 4391 | 5939 | 3068 | 0.530 | 580 |
| 12 IL | 11251 | 7.50 | 5903 | 5126 | 14186 | 5301 | 0.525 | 471 |
| 13 MI | 9082 | 7.00 | 5213 | 4817 | 6930 | 4768 | 0.574 | 525 |
| 14 WI | 4520 | 7.00 | 2465 | 4207 | 6580 | 2294 | 0.545 | 508 |
| 15 MN | 3896 | 7.00 | 2368 | 4332 | 8159 | 2204 | 0.608 | 566 |
| 16 IA | 2883 | 7.00 | 1689 | 4318 | 10340 | 1830 | 0.586 | 635 |
| 17 MO | 4753 | 7.00 | 2719 | 4206 | 8508 | 2865 | 0.572 | 603 |
| 18 ND | 632 | 7.00 | 341 | 3718 | 4725 | 451 | 0.540 | 714 |
| 19 SD | 579 | 7.00 | 419 | 4716 | 5915 | 501 | 0.724 | 865 |
| 20 NE | 1525 | 8.50 | 1033 | 4341 | 6010 | 976 | 0.677 | 640 |
| 21 KS | 2258 | 7.00 | 1496 | 4593 | 7834 | 1466 | 0.663 | 649 |
| 22 DE | 565 | 8.00 | 340 | 4983 | 602 | 305 | 0.602 | 540 |
| 23 MD | 4056 | 9.00 | 2073 | 4897 | 2449 | 1883 | 0.511 | 464 |
| 24 VA | 4764 | 9.00 | 2463 | 4258 | 4686 | 2604 | 0.517 | 547 |
| 25 WV | 1781 | 8.50 | 982 | 4574 | 2619 | 819 | 0.551 | 460 |
| 26 NC | 5214 | 9.00 | 2835 | 3721 | 4746 | 2953 | 0.544 | 566 |
| 27 SC | 2665 | 8.00 | 1460 | 3448 | 5399 | 1537 | 0.548 | 577 |
| 28 GA | 4720 | 7.50 | 2731 | 3846 | 9061 | 2979 | 0.579 | 631 |
| 29 FA | 7259 | 8.00 | 4084 | 4188 | 5975 | 4169 | 0.563 | 574 |
| 30 KY | 3299 | 9.00 | 1626 | 3601 | 4650 | 1761 | 0.493 | 534 |
| 31 TN | 4031 | 7.00 | 2088 | 3640 | 6905 | 2301 | 0.518 | 571 |
| 32 AL | 3510 | 7.00 | 1801 | 3333 | 6594 | 1946 | 0.513 | 554 |
| 33 MS | 2263 | 8.00 | 1309 | 3063 | 6524 | 1306 | 0.578 | 577 |
| 34 AR | 1978 | 7.50 | 1081 | 3357 | 4121 | 1242 | 0.547 | 628 |
| 35 LA | 3720 | 8.00 | 1813 | 3528 | 3495 | 1812 | 0.487 | 487 |
| 36 OK | 2634 | 6.58 | 1657 | 3802 | 7834 | 1695 | 0.629 | 644 |
| 37 TX | 11649 | 5.00 | 6595 | 4045 | 17782 | 7451 | 0.566 | 640 |
| 38 MT | 719 | 7.00 | 421 | 3897 | 6385 | 506 | 0.586 | 704 |

## Statistical Meaning of Multiple Regression

$$\widehat{FUEL} = \hat{\beta}_0 + \hat{\beta}_1 TAX + \hat{\beta}_2 DLIC$$

?

?

$$\widehat{FUEL} = 984.0 - 53.11 TAX$$

## Removing Effects of DLIC

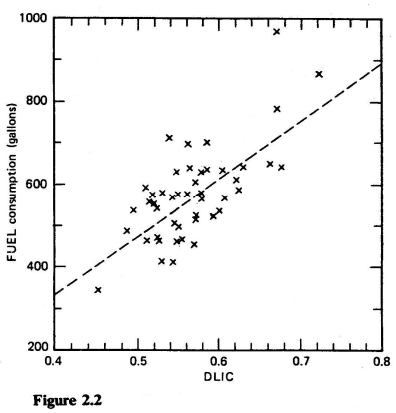



Figure 2.2

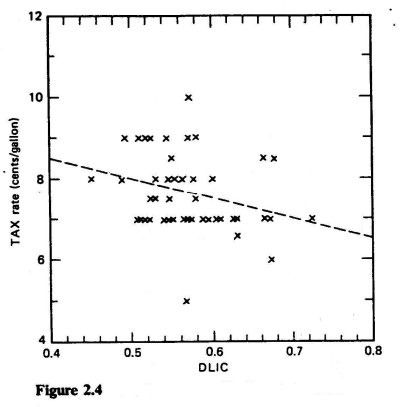



Figure 2.4

$$\widehat{FUEL} = 22.73 + 1409.8 DLIC$$

$$\widehat{TAX} = 10.48 - 4.94 DLIC$$

# True Effect of TAX

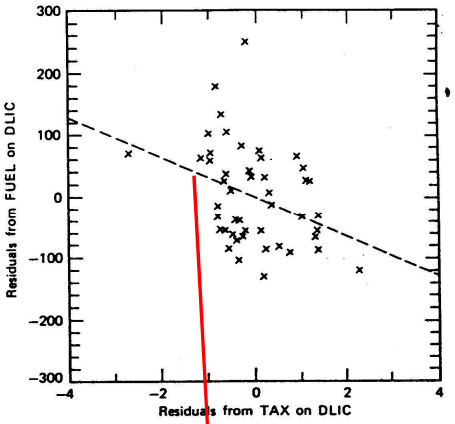



Figure 2.5

$$\widehat{FUEL} = 109.0 - 32.08TAX + 1251.5DLIC$$

# In Most Designs of Experiments

- X's are designed to be orthogonal (statistically uncorrelated) to one another
- $L_9$ $(3^4)$

| Expt. No. | Column 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 2 | 2 | 2 |
| 3 | 1 | 3 | 3 | 3 |
| 4 | 2 | 1 | 2 | 3 |
| 5 | 2 | 2 | 3 | 1 |
| 6 | 2 | 3 | 1 | 2 |
| 7 | 3 | 1 | 3 | 2 |
| 8 | 3 | 2 | 1 | 3 |
| 9 | 3 | 3 | 2 | 1 |

- $2^3$ Factorial

| | | Main Effects | | | Interactions | | | |
|---|---|---|---|---|---|---|---|---|
| Test | $I$ | $x_1$ | $x_2$ | $x_3$ | $x_1x_2$ | $x_1x_3$ | $x_2x_3$ | $x_1x_2x_3$ |
| 1 | + | -1 | -1 | -1 | +1 | +1 | +1 | -1 |
| 2 | + | +1 | -1 | -1 | -1 | -1 | +1 | +1 |
| 3 | + | -1 | +1 | -1 | -1 | +1 | -1 | +1 |
| 4 | + | +1 | +1 | -1 | +1 | -1 | -1 | -1 |
| 5 | + | -1 | -1 | +1 | +1 | -1 | -1 | +1 |
| 6 | + | +1 | -1 | +1 | -1 | +1 | -1 | -1 |
| 7 | + | -1 | +1 | +1 | -1 | -1 | +1 | -1 |
| 8 | + | +1 | +1 | +1 | +1 | +1 | +1 | +1 |

32

# With Uncorrelated Variables X's

- Back to the example of the Glove Box $2^4$ factorial design of experiments

  Multiple Regression $b_i$ = Simple Regression $b_i$

| | 係數 | 標準誤 | t 統計 | P-值 | 下限 95% | 上限 95% | 下限 95.0% | 上限 95.0% |
|---|---|---|---|---|---|---|---|---|
| 截距 | -0.08719 | 0.079537 | -1.09619 | 0.28922 | -0.2557976 | 0.08142257 | -0.2557976 | 0.08142257 |
| X1 | -0.32719 | 0.079537 | -4.11367 | 0.000813 | -0.4957976 | -0.1585774 | -0.4957976 | -0.15857743 |
| X2 | 0.397188 | 0.079537 | 4.993769 | 0.000133 | 0.22857743 | 0.56579757 | 0.22857743 | 0.56579757 |
| X3 | 0.319063 | 0.079537 | 4.011517 | 0.001007 | 0.15045243 | 0.48767257 | 0.15045243 | 0.48767257 |
| X4 | 0.160938 | 0.079537 | 2.023439 | 0.060064 | -0.0076726 | 0.32954757 | -0.0076726 | 0.32954757 |
| X1X2 | 0.073438 | 0.079537 | 0.923317 | 0.369557 | -0.0951726 | 0.24204757 | -0.0951726 | 0.24204757 |
| X1X3 | -0.05844 | 0.079537 | -0.73472 | 0.473139 | -0.2270476 | 0.11017257 | -0.2270476 | 0.11017257 |
| X1X4 | -0.01531 | 0.079537 | -0.19252 | 0.849756 | -0.1839226 | 0.15329757 | -0.1839226 | 0.15329757 |
| X2X3 | -0.09531 | 0.079537 | -1.19835 | 0.248231 | -0.2639226 | 0.07329757 | -0.2639226 | 0.07329757 |
| X2X4 | -0.07719 | 0.079537 | -0.97046 | 0.346258 | -0.2457976 | 0.09142257 | -0.2457976 | 0.09142257 |
| X3X4 | 0.004688 | 0.079537 | 0.058935 | 0.953734 | -0.1639226 | 0.17329757 | -0.1639226 | 0.17329757 |
| X1X2X3 | 0.085938 | 0.079537 | 1.080477 | 0.295948 | -0.0826726 | 0.25454757 | -0.0826726 | 0.25454757 |
| X1X2X4 | 0.050313 | 0.079537 | 0.63257 | 0.535951 | -0.1182976 | 0.21892257 | -0.1182976 | 0.21892257 |
| X1X3X4 | -0.06906 | 0.079537 | -0.86831 | 0.398062 | -0.2376726 | 0.09954757 | -0.2376726 | 0.09954757 |
| X2X3X4 | -0.05219 | 0.079537 | -0.65614 | 0.521056 | -0.2207976 | 0.11642257 | -0.2207976 | 0.11642257 |
| X1X2X3X | 0.060313 | 0.079537 | 0.758298 | 0.459297 | -0.1082976 | 0.22892257 | -0.1082976 | 0.22892257 |

| | *係數* | *標準誤* | *t 統計* | *P-值* | *下限 95%* | *上限 95%* | *下限 95.0%* | *上限 95.0%* |
|---|---|---|---|---|---|---|---|---|
| 截距 | -0.08719 | 0.112713 | -0.77354 | 0.445257 | -0.31738 | 0.143003 | -0.31738 | 0.143003 |
| X2 | 0.397188 | 0.112713 | 3.523892 | 0.001386 | 0.166997 | 0.627378 | 0.166997 | 0.627378 |

33

# Multicollinearity in Linear Regression

- Example: House Price
  - A real estate agent believes that a house selling price can be predicted using the house size, number of bedrooms, and lot size.
  - A random sample of 100 houses was drawn and data recorded.

| Price | Bedrooms | H Size | Lot Size |
|---|---|---|---|
| 124100 | 3 | 1290 | 3900 |
| 218300 | 4 | 2080 | 6600 |
| 117800 | 3 | 1250 | 3750 |
| . | . | . | . |
| . | . | . | . |

  - Analyze the relationship among the four variables

- Solution
- The proposed model is
  **PRICE =** $\beta_0$ + $\beta_1$**BEDROOMS** + $\beta_2$**H-SIZE** +$\beta_3$**LOTSIZE** + $\varepsilon$
  - Excel solution

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.74833 |
| R Square | 0.559998 |
| Adjusted F | 0.546248 |
| Standard E | 25022.71 |
| Observatio | 100 |

The model is valid, but no variable is significantly related to the selling price !!

ANOVA

| | *df* | *SS* | *MS* | *F* | *'gnificance F* |
|---|---|---|---|---|---|
| Regressior | 3 | 7.65E+10 | 2.55E+10 | 40.7269 | 4.57E-17 |
| Residual | 96 | 6.01E+10 | 6.26E+08 | | |
| Total | 99 | 1.37E+11 | | | |

| | *Coefficients* | *tandard Em* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 37717.59 | 14176.74 | 2.660526 | 0.009145 | 9576.963 | 65858.23 |
| Bedrooms | 2306.081 | 6994.192 | 0.329714 | 0.742335 | -11577.3 | 16189.45 |
| H Size | 74.29681 | 52.97858 | 1.402393 | 0.164023 | -30.8649 | 179.4585 |
| Lot Size | -4.36378 | 17.024 | -0.25633 | 0.798244 | -38.1562 | 29.42862 |

- However,
  - when regressing the price on each independent variable alone, it is found that each variable is strongly related to the selling price.
  - Multicollinearity is the source of this problem.

| | ***Price*** | ***Bedrooms*** | ***H Size*** | ***Lot Size*** |
|---|---|---|---|---|
| Price | 1 | | | |
| Bedrooms | 0.645411 | 1 | | |
| H Size | 0.747762 | 0.846454 | 1 | |
| Lot Size | 0.740874 | 0.83743 | 0.993615 | 1 |

- Multicollinearity causes two kinds of difficulties:
  - The t statistics appear to be too small.
  - The $\beta$ coefficients cannot be interpreted as "slopes".

## Nonlinearity in Linear Regression

- Regression analysis is considered powerful for several reasons:
  - It can cover variety of mathematical models
    - linear relationships.
    - non - linear relationships.
    - qualitative variables.
  - It provides efficient methods for model building, to select the best fitting set of variables.

## Polynomial Models

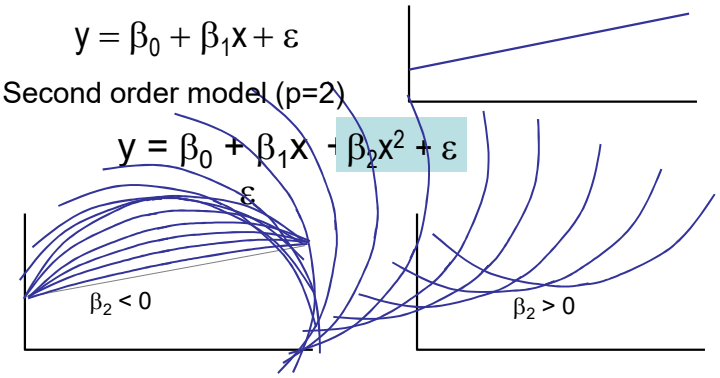
- The independent variables may appear as functions of a number of predictor variables.
  - Polynomial models of order p with one predictor variable: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p + \varepsilon$
  - Polynomial models with two predictor variables
    For example:
    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

- Polynomial models with one predictor variable
  - First order model (p = 1)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

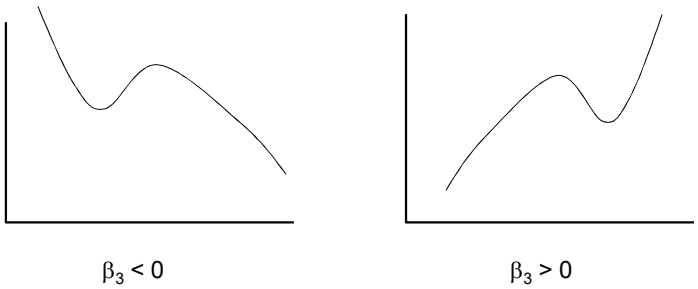
  - Second order model (p=2)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

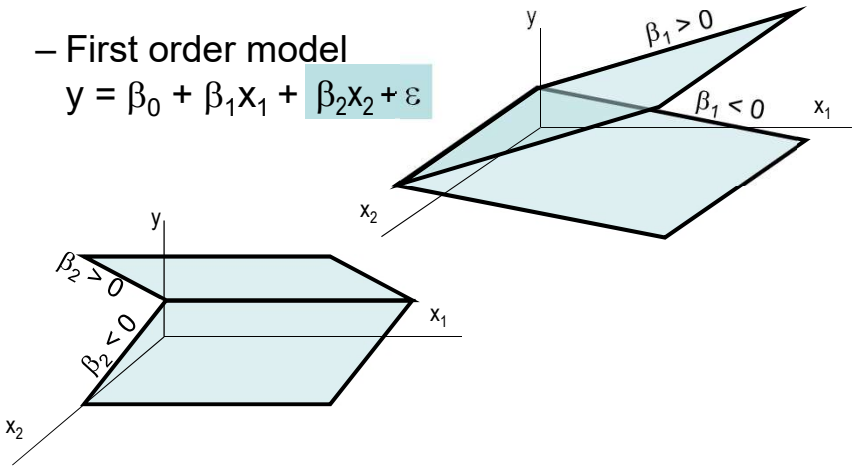



  - Third order model (p=3)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

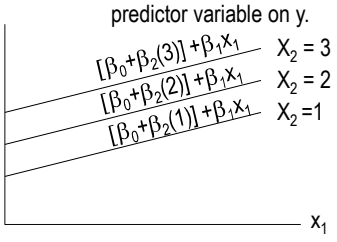


- Polynomial models with two predictor variables
  - First order model
    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



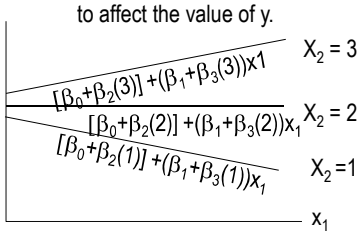

- Polynomial models with two predictor variables
  - First order model
    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

The effect of one predictor variable on y is independent of the effect of the other predictor variable on y.



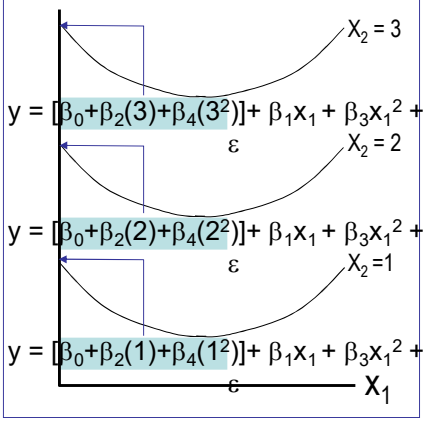

  - First order model with interaction
    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

The two variables interact to affect the value of y.



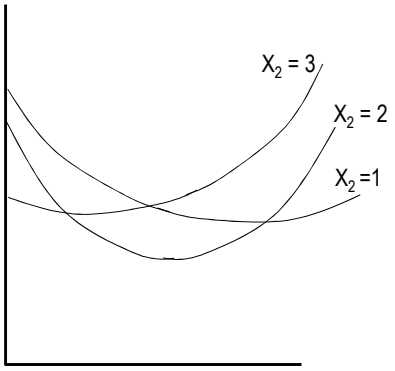
– Second order model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \varepsilon$



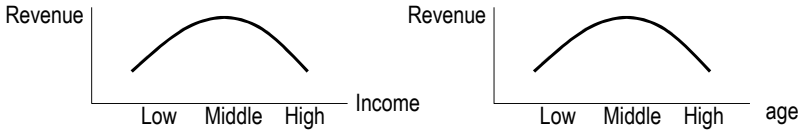

– Second order model with interaction

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$




- Example Location for a new restaurant
  - A fast food restaurant chain tries to identify new locations that are likely to be profitable.
  - The primary market for such restaurants is middle-income adults and their children (between the age 5 and 12).
  - Which regression model should be proposed to predict the profitability of new locations?

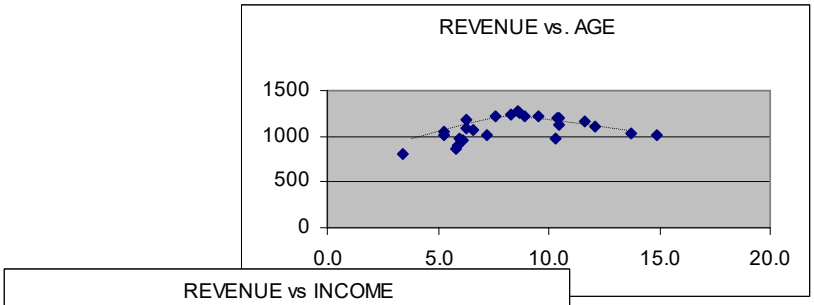
- Solution
  - The dependent variable will be **Gross Revenue**
  - There are quadratic relationships between Revenue and each predictor variable. Why?
    - Members of middle-class families are more likely to visit a fast food family than members of poor or wealthy families.

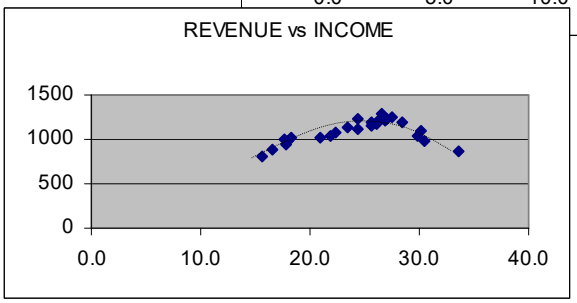



    - Families with very young or older kids will not visit the restaurant as frequent as families with mid-range ages of kids.

$$\textbf{Revenue} = \beta_0 + \beta_1\textbf{Income} + \beta_2\textbf{Age} + \beta_3\textbf{Income}^2 + \beta_4\textbf{Age}^2 + \beta_5(\textbf{Income})(\textbf{Age}) + \varepsilon$$

- Example
  - To verify the validity of the model proposed in example, 25 areas with fast food restaurants were randomly selected.
  - Data collected included (see Xm19-02.xls):
    - Previous year's annual gross sales.
    - Mean annual household income.
    - Mean age of children




SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.952121 |
| R Square | 0.906535 |
| Adjusted F | 0.881939 |
| Standard E | 44.69533 |
| Observatio | 25 |

The model provides a good fit

ANOVA

| | *df* | *SS* | *MS* | *F* | *ignificance F* |
|---|---|---|---|---|---|
| Regressior | 5 | 368140.4 | 73628.08 | 36.85692 | 3.86E-09 |
| Residual | 19 | 37955.78 | 1997.673 | | |
| Total | 24 | 406096.2 | | | |

| | *Coefficients* | *tandard Er* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | -1133.98 | 320.0193 | -3.54348 | 0.00217 | -1803.79 | -464.173 |
| INCOME | 173.2032 | 28.20399 | 6.141086 | 6.66E-06 | 114.1715 | 232.2348 |
| AGE | 23.54996 | 32.23447 | 0.730583 | 0.473947 | -43.9176 | 91.01751 |
| INC sq | -3.72613 | 0.542156 | -6.8728 | 1.48E-06 | -4.86087 | -2.59138 |
| AGE sq | -3.86871 | 1.179054 | -3.28119 | 0.003928 | -6.3365 | -1.40092 |
| INC X AGE | 1.967268 | 0.944082 | 2.08379 | 0.050921 | -0.00872 | 3.943255 |

The model can be used to make predictions.

However, do not interpret the coefficients or test them.
Multicollinearity is a problem!!

In excel: Tools > Data Analysis > Correlation

| | INCOME | AGE | INC sq | AGE sq | INC X AGE |
|---|---|---|---|---|---|
| INCOME | 1 | | | | |
| AGE | 0.0201 | 1 | | | |
| INC sq | 0.9945 | -0.045 | 1 | | |
| AGE sq | -0.042 | 0.9845 | -0.099 | 1 | |
| INC X AGE | 0.4596 | 0.8861 | 0.3968 | 0.8405 | 1 |

The multicollinearity can be reduced by modifying the original predictor variables

Income = **Income - average income**
Age = **Age - average Age**

Income = **Income - 24.2**

Age = **Age - 8.392**

Original data

| REVENUE | INCOME | AGE | INC sq | AGE sq | INC X AGE |
|---|---|---|---|---|---|
| 1128 | 23.5 | 10.5 | 552.25 | 110.25 | 246.75 |
| 1005 | 17.6 | 7.2 | 309.76 | 51.84 | 126.72 |
| 1212 | 26.3 | 7.6 | 691.69 | 57.76 | 199.88 |
| 893 | | | | | |
| 1073 | | | | | |
| 1179 | | | | | |
| 1109 | | | | | |
| 1019 | | | | | |
| 1228 | | | | | |

$(-.7)^2$

(-.7)(2.11)

| REVENUE | Income | Age | Inc-sq | Age sq | Inc X Age |
|---|---|---|---|---|---|
| 1128 | -0.7 | 2.11 | 0.49 | 4.4437 | -1.4756 |
| 1005 | -6.6 | -1.2 | 43.56 | 1.4209 | 7.8672 |
| 1212 | 2.1 | -0.8 | 4.41 | 0.6273 | -1.6632 |
| . | . | . | . | . | . |
| . | . | . | . | . | . |

Modified data

# Regression results of the modified model

SUMMARY OUTPUT

Multicolinearity is not a problem anymore

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.952121 |
| R Square | 0.906535 |
| Adjusted R Sq | 0.881939 |
| Standard Error | 44.69533 |
| Observations | 25 |

| | *Income* | *Age* | *Inc-sq* | *Age sq* | *Inc X Age* |
|---|---|---|---|---|---|
| Income | 1 | | | | |
| Age | 0.020058 | 1 | | | |
| Inc-sq | -0.17263 | -0.61236 | 1 | | |
| Age sq | -0.31219 | 0.400929 | 0.096296 | 1 | |
| Inc X Age | -0.61438 | -0.37221 | 0.167909 | -0.08263 | 1 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 5 | 368140.4 | 73628.08 | 36.85692 | 3.86E-09 |
| Residual | 19 | 37955.78 | 1997.673 | | |
| Total | 24 | 406096.2 | | | |

| | *Coefficients* | *Standard Err* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 1200.066 | 15.08728 | 79.54156 | 1.91E-25 | 1168.488 | 1231.644 |
| Income | 9.367849 | 2.743887 | 3.41408 | 0.00291 | 3.624827 | 15.11087 |
| Age | 6.225473 | 5.472777 | 1.137534 | 0.269457 | -5.229186 | 17.68013 |
| Income sq | -3.72613 | 0.542156 | -6.8728 | 1.48E-06 | -4.860874 | -2.59138 |
| Age sq | -3.86871 | 1.179054 | -3.28119 | 0.003928 | -6.336497 | -1.40092 |
| Inc X Age | 1.967268 | 0.944082 | 2.08379 | 0.050921 | -0.008718 | 3.943255 |