Regression Approach to ANOVA

Design of Experiments - Montgomery
Section 3-9, Chapter 10

The Regression Approach

One-way Anova

- Consider the ANOVA model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots n_i \end{cases}$$

- Will compare it to the regression model

$$y_j = \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_a x_{a,j} + \epsilon_j$$

where

$$j = 1, 2, \ldots N = \sum n_i$$

and, for example,

$$x_{i,j} \;=\; \begin{cases} 1 \text{ if } j\text{th obs is from trt } i \\ 0 \text{ otherwise} \end{cases}$$

- Will use β's for regression, μ and τ's for ANOVA

Fitting Regression Models

- Can write the regression model as $y = X\beta + \epsilon$ where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_a \end{pmatrix} \quad X = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{a,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{a,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,N} & x_{2,N} & \cdots & x_{a,N} \end{bmatrix}$$

- Least squares estimator of β is $\hat{\beta} = (X'X)^{-1}X'y$
- Least squares normal equations are $(X'X)\hat{\beta} = X'y$
- The vector of predicted values $\hat{y} = X\hat{\beta}$
- The vector of residual values $\hat{\epsilon} = y - \hat{y} = y - X\hat{\beta}$
- $\text{SS}_\text{E} = \sum_{j=1}^{N} \hat{\epsilon}_j^2 = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - (X\hat{\beta})'y$
- The estimated covariance matrix is $\text{Cov}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$
- To estimate a linear combination C of the β parameters, use

$$C'\hat{\beta} \qquad \text{Var}(C'\hat{\beta}) = \sigma^2 C'(X'X)^{-1}C$$

General Regression Significance Test

- Suppose $\boldsymbol{\beta} = (\boldsymbol{\beta_A}, \boldsymbol{\beta_B})'$

  For example, $\boldsymbol{\beta_A} = \beta_0$ and $\boldsymbol{\beta_B} = (\beta_1, ..., \beta_a)$

- Want to test $H_0 : \boldsymbol{\beta_B} = 0$
  - Define $\boldsymbol{\beta}^\star = (\boldsymbol{\beta_A}, 0)'$
  - Fit $y = X\boldsymbol{\beta}$ (Full) and $y = X\boldsymbol{\beta}^\star$ (Reduced)

$$F_0 = \frac{(\text{SS}_\text{E}(\text{Reduced}) - \text{SS}_\text{E}(\text{Full}))/(\text{df}_\text{R} - \text{df}_\text{F})}{\text{SS}_\text{E}(\text{Full})/\text{df}_\text{F}}$$

  - If $\epsilon \sim N(0, \sigma^2)$ and H_0 true, $F_0 \sim F_{\text{df}_R - \text{df}_F, \text{df}_F}$

- Can generate F test for any subset of $\boldsymbol{\beta}$

Orthogonality

- Recall orthogonal contrasts
  - If C and D contrasts, $\sum c_i d_i / n_i = 0$
  - If a treatments, can construct $a-1$ orthogonal contrasts
  - SS$_{\text{Trt}}$ separated in $a-1$ indep SS with 1 df
  - Order of computation does not affect results
- Orthogonal regression variables
  - Regressors x_1 and x_2 are orthogonal if $\sum x_{1j} x_{2j} = 0$
  - Order of partitioning does not affect SS results
  - Experimental designs form sets of orthogonal regressors

Application to CRD Design

Consider an experiment with $n = 3$, $a = 3$ and data set:

Trt1	Trt 2	Trt 3
20	24	26
23	26	27
22	25	27

$\overline{y}_{..} = 24.44 \quad \overline{y}_{1.} = 21.67 \quad \overline{y}_{2.} = 25.00 \quad \overline{y}_{3.} = 26.67$

- SAS converts GLM commands into matrices
- Includes column for intercept and **ALL** treatments
- Defining $x_{i,j}$ as in slide 9-1

$$Y = \begin{bmatrix} 20 \\ 23 \\ 22 \\ 24 \\ 26 \\ 25 \\ 26 \\ 27 \\ 27 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad X'X = \begin{bmatrix} 9 & 3 & 3 & 3 \\ 3 & 3 & 0 & 0 \\ 3 & 0 & 3 & 0 \\ 3 & 0 & 0 & 3 \end{bmatrix}$$

- Over-parameterized model (rank $X'X = a$)
- Uses generalized inverse (similar to setting $\beta_3 = 0$)

CRD Example Continued

- Setting $\beta_3 = 0 \rightarrow$ removing last column of X
- When observation from

  Trt 1 $\rightarrow E(y_j) = \beta_0 + \beta_1$

  Trt 2 $\rightarrow E(y_j) = \beta_0 + \beta_2$

  Trt 3 $\rightarrow E(y_j) = \beta_0$
- Can show the linear regression estimates are

  $\hat{\beta}_0 = \overline{y}_{3.}$

  $\hat{\beta}_1 = \overline{y}_{1.} - \overline{y}_{3.}$

  $\hat{\beta}_2 = \overline{y}_{2.} - \overline{y}_{3.}$
- In terms of CRD model parameters

  $\beta_0 = \mu + \tau_3$

  $\beta_1 = \tau_1 - \tau_3$

  $\beta_2 = \tau_2 - \tau_3$

CRD Example Continued

- Now consider the constraint $\sum_{j=1}^{a} \beta_j = 0$
- Since $\beta_3 = -\beta_1 - \beta_2$, can write X as

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \qquad X'X = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 6 & 3 \\ 0 & 3 & 6 \end{bmatrix}$$

- When observation from

  Trt 1 $\rightarrow E(y_j) = \beta_0 + \beta_1$

  Trt 2 $\rightarrow E(y_j) = \beta_0 + \beta_2$

  Trt 3 $\rightarrow E(y_j) = \beta_0 - \beta_1 - \beta_2$
- The parameter estimates are

  $\hat{\beta}_0 = \overline{y}_{..} \qquad \rightarrow \beta_0 = \mu$

  $\hat{\beta}_1 = \overline{y}_{1.} - \overline{y}_{..} \qquad \rightarrow \beta_1 = \tau_1$

  $\hat{\beta}_2 = \overline{y}_{2.} - \overline{y}_{..} \qquad \rightarrow \beta_2 = \tau_2$

Using SAS

```
option nocenter ps=50 ls=72;

data one;
 input trt response @@;
 cards;
 1 20 1 23 1 22
 2 24 2 26 2 25
 3 26 3 27 3 27
;

data two;
 set one;
 if trt = 1 then x1=1;
 else x1=0;
 if trt = 2 then x2=1;
 else x2=0;

data three;
 set one;
 if trt = 1 then x1=1;
 else if trt = 3 then x1=-1;
 else x1=0;
 if trt = 2 then x2=1;
 else if trt = 3 then x2=-1;
 else x2=0;

proc glm data=one;
 class trt;
 model response=trt /solution xpx;

proc reg data=two;
 model response=x1 x2;

proc reg data=three;
 model response=x1 x2;
```

```
General Linear Models Procedure

The X'X Matrix
            INTERCEPT  TRT 1     TRT 2     TRT 3   RESPONSE
INTERCEPT      9         3         3         3       220
TRT 1          3         3         0         0        65
TRT 2          3         0         3         0        75
TRT 3          3         0         0         3        80
RESPONSE     220        65        75        80      5424

Dependent Variable: RESPONSE
                                   Sum of         Mean
Source                   DF       Squares       Square  F Value     Pr > F
Model                     2     38.888889    19.444444    15.91     0.0040
Error                     6      7.333333     1.222222
Corrected Total           8     46.222222

                 R-Square          C.V.      Root MSE      RESPONSE Mean
                 0.841346      4.522670        1.1055             24.444

Source                   DF     Type I SS  Mean Square  F Value     Pr > F
TRT                       2     38.888889    19.444444    15.91     0.0040

                                  T for H0:     Pr > |T|     Std Error of
Parameter           Estimate     Parameter=0                  Estimate
INTERCEPT       26.66666667 B         41.78       0.0001       0.63828474
TRT        1    -5.00000000 B         -5.54       0.0015       0.90267093
           2    -1.66666667 B         -1.85       0.1144       0.90267093
           3     0.00000000 B           .           .              .

NOTE: The X'X matrix has been found to be singular and a generalized
      inverse was used to solve the normal equations.   Estimates
      followed by the letter 'B' are biased, and are not unique
      estimators of the parameters.
```

```
______________________Model 1________________________________________
Analysis of Variance
                         Sum of         Mean
Source          DF      Squares       Square      F Value       Prob>F
Model            2     38.88889     19.44444       15.909       0.0040
Error            6      7.33333      1.22222
C Total          8     46.22222

    Root MSE       1.10554     R-square       0.8413
    Dep Mean      24.44444     Adj R-sq       0.7885
    C.V.           4.52267

Parameter Estimates
                 Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|
INTERCEP   1     26.666667    0.63828474        41.779        0.0001
X1         1     -5.000000    0.90267093        -5.539        0.0015
X2         1     -1.666667    0.90267093        -1.846        0.1144

____________________Model 2__________________________________________

Analysis of Variance
                         Sum of         Mean
Source          DF      Squares       Square      F Value       Prob>F
Model            2     38.88889     19.44444       15.909       0.0040
Error            6      7.33333      1.22222
C Total          8     46.22222

    Root MSE       1.10554     R-square       0.8413
    Dep Mean      24.44444     Adj R-sq       0.7885
    C.V.           4.52267

Parameter Estimates
                 Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|
INTERCEP   1     24.444444    0.36851387        66.332        0.0001
X1         1     -2.777778    0.52115731        -5.330        0.0018
X2         1      0.555556    0.52115731         1.066        0.3274
```

Summary of Two Models

- In both cases $H_0 : \beta_1 = \beta_2 = 0$ similar to usual H_0
- Model #1 : Similar to SAS

$$\widehat{\beta}_0 = \overline{y}_{3.},\ \widehat{\beta}_1 = \overline{y}_{1.} - \overline{y}_{3.},\ \widehat{\beta}_2 = \overline{y}_{2.} - \overline{y}_{3.}$$

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} 1/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & 1/3 \\ -1/3 & 1/3 & 2/3 \end{bmatrix}$$

- Model #2

$$\widehat{\beta}_0 = \overline{y}_{..},\ \widehat{\beta}_1 = \overline{y}_{1.} - \overline{y}_{..},\ \widehat{\beta}_2 = \overline{y}_{2.} - \overline{y}_{..}$$

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} 1/9 & 0 & 0 \\ 0 & 2/9 & -1/9 \\ 0 & -1/9 & 2/9 \end{bmatrix}$$

Estimable Functions

- For some comb's of β's, same answer across models
- Here are several examples based on previous models

$$\beta_0 + \beta_1 = \mu + \tau_1$$
$$\beta_0 + \beta_2 = \mu + \tau_2$$
$$\beta_1 - \beta_2 = \tau_1 - \tau_2$$
$$\beta_1 - \beta_3 = \tau_1 - \tau_3$$

- These unique comb's known as estimable functions

Can uniquely estimate trt differences

Can uniquely estimate trt means

Cannot uniquely estimate trt effects, grand mean

(Like Model 1) : $\mu = 6,\ \tau_1 = 3, \tau_2 = -6, \tau_3 = 0$

(Like Model 2) : $\mu = 5,\ \tau_1 = 4, \tau_2 = -5, \tau_3 = 1$

Estimable Functions in SAS

```
option nocenter ps=50 ls=72;

data one;
 input trt response @@;
 cards;
 1 20 1 23 1 22
 2 24 2 26 2 25
 3 26 3 27 3 27
;

proc glm data=one;
 class trt;
 model response=trt / e;

-----------------------------------------------------------------------

Class    Levels    Values
TRT           3    1 2 3

Number of observations in data set = 9


General Linear Models Procedure
General Form of Estimable Functions

Effect        Coefficients

INTERCEPT     L1

TRT       1   L2
          2   L3
          3   L1-L2-L3
```

Examples

- Consider $C = \beta_0 + 2\beta_1 - \beta_3$
  - Since $-1 = 1 - 2 - 0$, it is estimable
  - Model 1 : $(\mu + \tau_3) + 2(\tau_1 - \tau_3) - 0$
  - Model 2 : $\mu + 2\tau_1 - \tau_3$
- Consider $C = \beta_0 + 2\beta_1 + \beta_2 - \beta_3$
  - Since $-1 \neq 1 - 2 - 1$, it is not estimable
  - Model 1 : $(\mu + \tau_3) + 2(\tau_1 - \tau_3) + (\tau_2 - \tau_3) - 0$
  - Model 2 : $\mu + 2\tau_1 + \tau_2 - \tau_3$