Statistics 514: Design of Experiments

Topic 9

Topic Overview

This topic will cover

- Analysis of Covariance
- Optimal Designs
- Regression Issues and Estimable Functions
- Regression and Non-orthogonal Designs

Analysis of Covariance (Section 15-3)

Background

- Consider factor x which is correlated with y
- Can measure x but can't control/predict it (as with blocks)
- Nuisance factor x called a *covariate*
- ANCOVA adjusts y for effect of covariate x ("retrospective adjustment for bias")
- Combination of regression and analysis of variance
- Without adjustment, effects of x may
  - inflate σ^2
  - alter treatment comparison

Examples

- **Pretest/Posttest score analysis:** The gain in score y may be associated with the pretest score x. Analysis of covariance provides a way to "handicap" each student. That way, one does not need to find a group of students with similar pretest scores and randomly assign them to a control and treatment group. Similar to analyzing difference in scores.
- **Weight gain experiments in animals:** If wishing to compare different feeds, the weight gain y may be associated with the original weight of the animal. Analysis of covariance provides a way to use a herd and adjust for the varying original weights.

- **Comparing competing drug products:** The effect of the drug y after two hours (measured on a scale from 1 to 10) may be associated with the initial mental and physical shape of the subject. Variables describing the initial mental and physical shape may be used as covariates.

Model Description

- Consider single covariate in CRD
- Statistical model is

$$y_{i,j} = \mu + \tau_i + \beta(x_{i,j} - \bar{x}_{..}) + \epsilon_{i,j} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n_i \end{cases}$$

- Additional assumptions
  - $x_{i,j}$ not affected by treatment
  - x and y are linearly related
  - Constant slope
- General Procedure:
  - Fit one-way model (`y = trt`)
  - Fit one-way model (`x = trt`)
  - Regress residuals (`residuals1 = residuals2`)
  - Model estimates are

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\beta} &= E_{xy}/E_{xx} = \sum\sum(y_{i,j} - \bar{y}_{i.})(x_{i,j} - \bar{x}_{i.})/\sum\sum(x_{i,j} - \bar{x}_{i.})^2 \\ \hat{\tau}_i &= \bar{y}_{i.} - \bar{y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \end{aligned}$$

Analysis of Covariance

- Test $\text{H}_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$
  - Compare treatment means after adjusting for differences among treatments due to differences in covariate levels.
  - Treatment and covariate not orthogonal (order of fit matters)

$$F_0 = \frac{SS(\text{trt}|x)/a - 1}{SS_E/(N - a - 1)}$$

- Adjusted treatment means
  - Estimate $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..})$

  - Variance: $\hat{\sigma}^2(1/n + (\bar{x}_{i.} - \bar{x}_{..})^2/\sum\sum(x_{i,j} - \bar{x}_{i.})^2)$

Could test $\beta = 0$

Sum of Squares regression (SS_x): $\hat{\beta}^2 \sum\sum(x_{i,j} - \bar{x}_{i.})^2$

$$F_0 = \frac{SS_x/1}{SS_E/(N-a-1)}$$

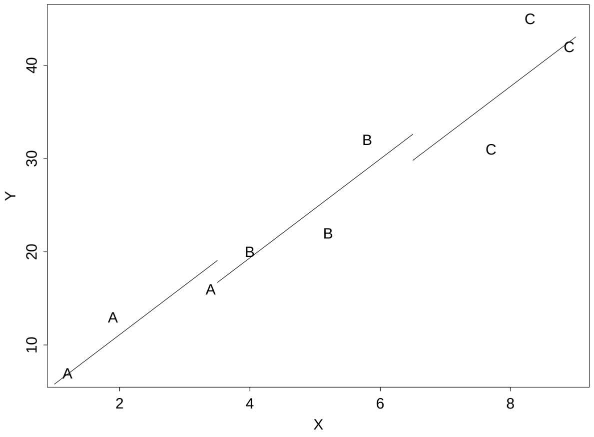
Relative Efficiency?

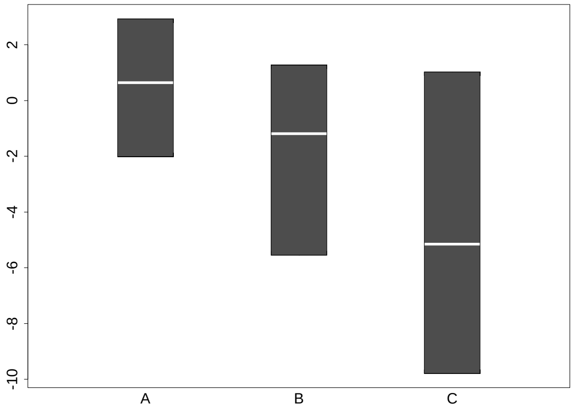
- Never used in literature
- Covariate not a *design* issue
- Small loss in df
- Makes assumptions about domain of x
- Not worried about number of covariates (as opposed to more complicated relationships)
- Using covariate might *negate* assumptions

Analysis of Covariance

Two Examples

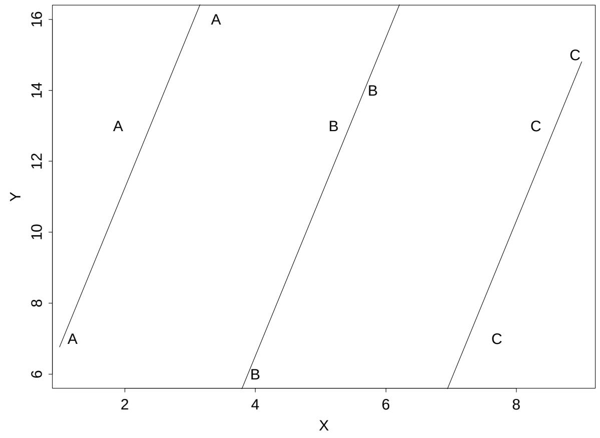
1. No treatment differences
   - Positive linear relationship
   - Covariate larger in each group
   - Thus, appears to be treatment difference ("failure of randomization")

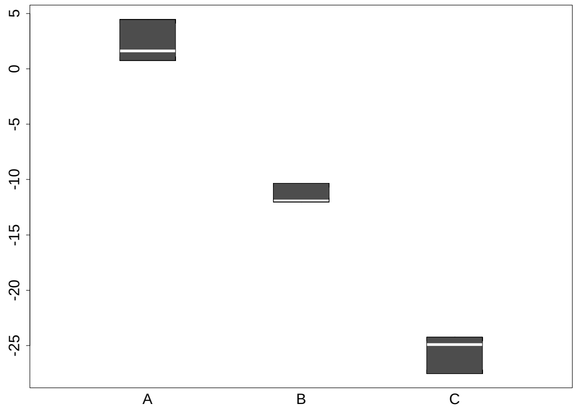




2. Treatment differences exist

- Positive linear relationship
- Covariate trend in each group
- Thus, no apparent treatment difference

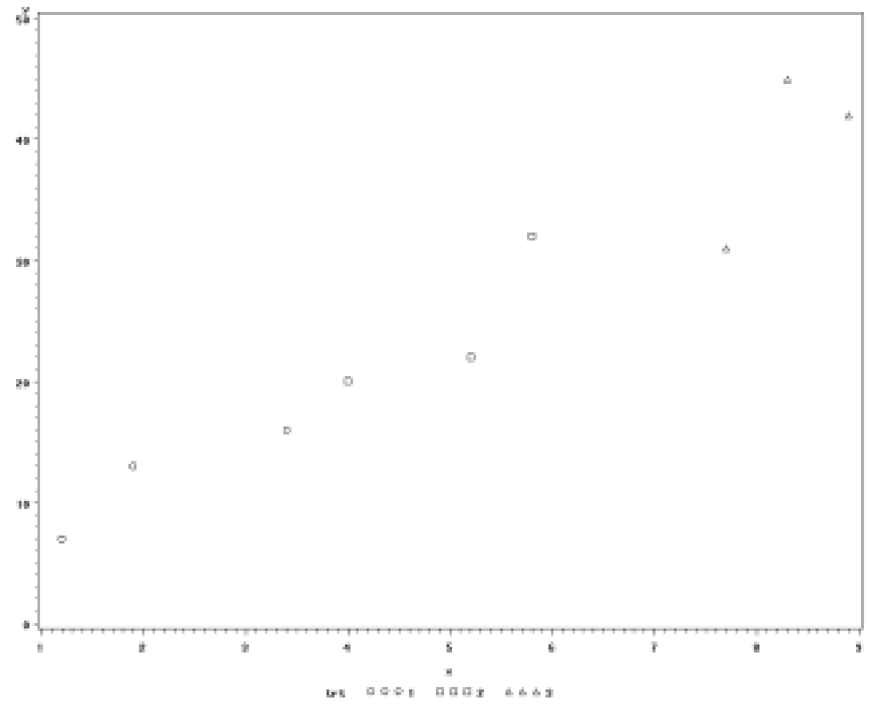




Using SAS

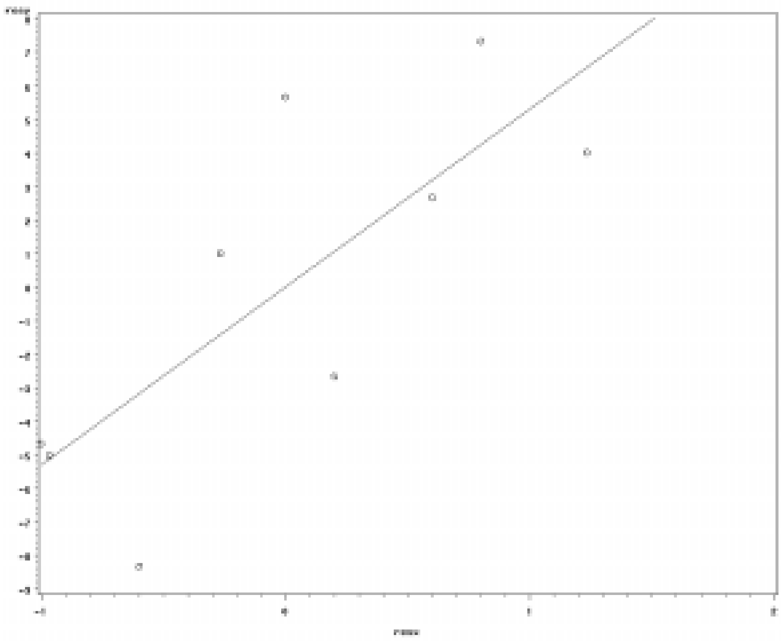
```
options nocenter ls=80;

data example1;
  input trt x y @@;
  cards;
  1 1.2 7 1 1.9 13 1 3.4 16
  2 4.0 20 2 5.2 22 2 5.8 32
  3 7.7 31 3 8.3 45 3 8.9 42
  ;

proc sort; by trt;
symbol1 v=circle   i=  c=black;
symbol2 v=square   i=  c=black;
symbol3 v=triangle i=  c=black;
proc gplot;
  plot y*x=trt;
run;

proc glm; class trt;
 model y=trt;
 output out=resid r=resy;
proc glm; class trt;
 model x=trt;
 output out=resid1 r=resx;
proc glm; model resy=resx;
symbol1 v=circle i=rl;
```

```
proc gplot;
  plot resy*resx;
run;

proc glm data=example1;
 class trt;
 model y=trt x / solution;
 means trt /lines lsd;
 lsmeans trt / tdiff adjust=t;
run;
```



```
The GLM Procedure
Dependent Variable: resy
                                      Sum of
Source                      DF       Squares    Mean Square   F Value   Pr > F
Model                        1   138.2699594    138.2699594     10.18   0.0153
Error                        7    95.0633739     13.5804820
Corrected Total              8   233.3333333

R-Square     Coeff Var      Root MSE     resy Mean
0.592586    1.03728E17      3.685171    3.5527E-15

Source                      DF     Type I SS    Mean Square   F Value   Pr > F
resx                         1   138.2699594    138.2699594     10.18   0.0153

Source                      DF   Type III SS    Mean Square   F Value   Pr > F
resx                         1   138.2699594    138.2699594     10.18   0.0153
```

```
                                        Standard
Parameter          Estimate                Error    t Value    Pr > |t|
Intercept       0.000000000           1.22839018       0.00      1.0000
resx            5.297699594           1.66027872       3.19      0.0153
```



```
The GLM Procedure
Dependent Variable: y
                                            Sum of
Source                        DF           Squares    Mean Square    F Value    Pr > F
Model                          3       1260.936626     420.312209      22.11    0.0026
Error                          5         95.063374      19.012675
Corrected Total                8       1356.000000

R-Square     Coeff Var      Root MSE        y Mean
0.929894      17.21192      4.360353      25.33333

Source                        DF         Type I SS    Mean Square    F Value    Pr > F
trt                            2       1122.666667     561.333333      29.52    0.0017
x                              1        138.269959     138.269959       7.27    0.0430

Source                        DF       Type III SS    Mean Square    F Value    Pr > F
trt                            2         3.2122606      1.6061303       0.08    0.9203
x                              1       138.2699594    138.2699594       7.27    0.0430

                                            Standard
Parameter            Estimate                  Error    t Value    Pr > |t|
Intercept       -4.637573297 B           16.49828508      -0.28      0.7899
trt       1      5.159224177 B           12.56372645       0.41      0.6983
trt       2      2.815741994 B            7.39601943       0.38      0.7191
trt       3      0.000000000 B             .                .         .
x                5.297699594              1.96446828       2.70      0.0430
```

```
t Tests (LSD) for y

Alpha                          0.05
Error Degrees of Freedom          5
Error Mean Square          19.01267
Critical Value of t         2.57058
Least Significant Difference   9.1518

Means with the same letter are not significantly different.

           Mean      N    trt
A        39.333      3    3
B        24.667      3    2
C        12.000      3    1


The GLM Procedure
Least Squares Means

                          LSMEAN
trt        y LSMEAN      Number
1        27.8342355           1
2        25.4907533           2
3        22.6750113           3


      Least Squares Means for Effect trt
   t for H0: LSMean(i)=LSMean(j) / Pr > |t|
            Dependent Variable: y
i/j              1             2             3

   1                    0.354685      0.410644
                          0.7373        0.6983
   2      -0.35468                     0.38071
            0.7373                      0.7191
   3      -0.41064      -0.38071
            0.6983        0.7191


NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

```
options nocenter ls=80;

data example2;
  input trt x y @@;
  cards;
  1 1.2 7 1 1.9 13 1 3.4 16
  2 4.0 6 2 5.2 13 2 5.8 14
  3 7.7 7 3 8.3 13 3 8.9 15
```

```
  ;

proc glm data = example2;
  class trt; model y = trt x / solution;
  lsmeans trt / tdiff;
  lsmeans trt / tdiff adjust = bon;
run;

                                              Sum of
Source                      DF           Squares     Mean Square    F Value    Pr > F
Model                        3       100.6915501      33.5638500      10.81    0.0126
Error                        5        15.5306721       3.1061344
Corrected Total              8       116.2222222

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
trt                          2      1.55555556      0.77777778       0.25    0.7877
x                            1     99.13599459     99.13599459      31.92    0.0024

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
trt                          2     94.55407736     47.27703868      15.22    0.0075
x                            1     99.13599459     99.13599459      31.92    0.0024

                                              Standard
Parameter           Estimate                 Error    t Value    Pr > |t|
Intercept       -25.56540370 B      6.66848712      -3.83      0.0122
trt       1      27.84618854 B      5.07816707       5.48      0.0028
trt       2      14.13644565 B      2.98941739       4.73      0.0052
trt       3       0.00000000 B       .                .          .
x                 4.48579161        0.79402382       5.65      0.0024


The GLM Procedure
Least Squares Means

trt        y LSMEAN      Number
1        25.4075327           1
2        11.6977898           2
3        -2.4386558           3

      Least Squares Means for Effect trt
   t for H0: LSMean(i)=LSMean(j) / Pr > |t|
            Dependent Variable: y
i/j              1             2             3
   1                    5.133597      5.483512
                          0.0037        0.0028
   2       -5.1336                     4.72883
            0.0037                      0.0052
   3      -5.48351      -4.72883
            0.0028        0.0052

Adjustment for Multiple Comparisons: Bonferroni
      Least Squares Means for Effect trt
   t for H0: LSMean(i)=LSMean(j) / Pr > |t|
```

```
            Dependent Variable: y

i/j              1            2            3

   1                   5.133597     5.483512
                         0.0110       0.0083
   2       -5.1336                   4.72883
            0.0110                    0.0156
   3      -5.48351     -4.72883
            0.0083       0.0156
```

Regression Approach to ANCOVA

- Consider the following model ($a = 3$)

$$y_j = \beta_0 + \beta_1 X_{1,j} + \beta_2 X_{2,j} + \beta_3 X_{3,j} + \epsilon_j$$

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

$$\begin{aligned} X_{1,j} &= \begin{cases} X_{1,j} = 1 & \text{if Treatment 1} \\ X_{1,j} = -1 & \text{if Treatment 3} \\ X_{1,j} = 0 & \text{otherwise} \end{cases} \\ X_{2,j} &= \begin{cases} X_{2,j} = 1 & \text{if Treatment 2} \\ X_{2,j} = -1 & \text{if Treatment 3} \\ X_{2,j} = 0 & \text{otherwise} \end{cases} \\ X_{3,j} &= (x_j - \bar{x}_{..}),\ x \text{ is the covariate} \end{aligned}$$

- Treatment 1: $y_j = \beta_0 + \beta_1 + \beta_3(x_j - \bar{x}_{..}) + \epsilon_j$
- Treatment 2: $y_j = \beta_0 + \beta_2 + \beta_3(x_j - \bar{x}_{..}) + \epsilon_j$
- Treatment 3: $y_j = \beta_0 - \beta_1 - \beta_2 + \beta_3(x_j - \bar{x}_{..}) + \epsilon_j$
- Results in estimates

$$\begin{array}{cc} \hat{\mu} = \hat{\beta}_0 & \hat{\tau}_1 = \hat{\beta}_1 \\ \hat{\tau}_2 = \hat{\beta}_2 & \hat{\beta} = \hat{\beta}_3 \end{array}$$

Nonconstant Slope in ANCOVA

- Statistical model for constant slope is

$$y_{i,j} = \mu + \tau_i + \beta(x_{i,j} - \bar{x}_{..}) + \epsilon_{i,j} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n_i \end{cases}$$

- Can allow for different slope by including interaction

$$y_{i,j} = \mu + \tau_i + (\beta - (\beta\tau)_i)(x_{i,j} - \bar{x}_{..}) + \epsilon_{i,j} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n_i \end{cases}$$

- In SAS, simply add interaction term into model
- Provides test for nonconstant slope

Using SAS

```
options nocenter ls=75;

data example1;
  input trt x y @@;
  cards;
  1 1.2 7 1 1.9 13 1 3.4 16
  2 4.0 20 2 5.2 22 2 5.8 32
  3 7.7 31 3 8.3 45 3 8.9 42
  ;

proc sort; by trt;
symbol1 v=circle   i=  c=black;
symbol2 v=square   i=  c=black;
symbol3 v=triangle i=  c=black;
proc gplot;
  plot y*x=trt;
run;

proc glm;
 class trt;
 model y=trt x / solution;
 lsmeans trt /tdiff;

proc glm;
 class trt;
 model y=trt x trt*x / solution;
 lsmeans trt / tdiff;
run;
```

```
                                      Sum of
Source                      DF       Squares   Mean Square  F Value  Pr > F
Model                        3   1260.936626    420.312209    22.11  0.0026
Error                        5     95.063374     19.012675
Corrected Total              8   1356.000000

Source                      DF   Type III SS   Mean Square  F Value  Pr > F
trt                          2     3.2122606     1.6061303     0.08  0.9203
```

```
x                          1   138.2699594   138.2699594     7.27  0.0430

                                      Standard
Parameter           Estimate             Error    t Value    Pr > |t|
Intercept      -4.637573297 B      16.49828508      -0.28      0.7899
trt       1     5.159224177 B      12.56372645       0.41      0.6983
trt       2     2.815741994 B       7.39601943       0.38      0.7191
trt       3     0.000000000 B        .                .         .
x               5.297699594         1.96446828       2.70      0.0430
--------------------------------------------------------------------------
                                    Sum of
Source                    DF       Squares   Mean Square  F Value  Pr > F
Model                      5   1278.409474    255.681895     9.89  0.0441
Error                      3     77.590526     25.863509
Corrected Total            8   1356.000000

Source                    DF   Type III SS   Mean Square  F Value  Pr > F
trt                        2    20.5146998    10.2573499     0.40  0.7034
x                          1   149.7599282   149.7599282     5.79  0.0953
x*trt                      2    17.4728475     8.7364237     0.34  0.7374

                                      Standard
Parameter           Estimate             Error    t Value    Pr > |t|
Intercept      -36.75000000 B      49.83227932      -0.74      0.5143
trt       1     40.60356201 B      50.39772400       0.81      0.4794
trt       2     31.65476190 B      53.63535098       0.59      0.5966
trt       3      0.00000000 B        .                .         .
x                9.16666667 B       5.99345810       1.53      0.2236
x*trt     1     -5.40677221 B       6.79395005      -0.80      0.4843
x*trt     2     -3.21428571 B       7.16355259      -0.45      0.6841
x*trt     3      0.00000000 B        .                .         .



                         LSMEAN
trt        y LSMEAN      Number
1        27.8342355           1
2        25.4907533           2
3        22.6750113           3

      Least Squares Means for Effect trt
   t for H0: LSMean(i)=LSMean(j) / Pr > |t|
            Dependent Variable: y
i/j              1             2             3
   1                    0.354685      0.410644
                          0.7373        0.6983
   2      -0.35468                     0.38071
            0.7373                      0.7191
   3      -0.41064      -0.38071
            0.6983        0.7191
------------------------------------------------
                         LSMEAN
trt        y LSMEAN      Number
```

```
1        23.2379068          1
2        25.5925926          2
3        10.5092593          3

      Least Squares Means for Effect trt
   t for H0: LSMean(i)=LSMean(j) / Pr > |t|
            Dependent Variable: y
i/j             1            2            3
   1                   -0.22548        0.591
                         0.8361       0.5961
   2      0.225476                  0.781205
            0.8361                    0.4917
   3        -0.591     -0.78121
            0.5961       0.4917
```

Analysis of Covariance

- Can incorporate covariate into any model
- For two-factor model

$$y_{i,j,k} = \mu + \tau_i + \beta_j + (\tau\beta)_{i,j} + \gamma(x_{i,j,k} - \bar{x}_{...}) + \epsilon_{i,j,k}$$

- Assume constant slope **for each** (i, j) **combination**
- Can include interaction terms to vary slope
- Plot y vs x for each combination

Optimal Designs

- So far, have not dealt with *assigning* points in continuous space
- In ANCOVA, points are recorded post-hoc.
- Could conceivably block/randomize treatment factors to keep balance/orthogonality
- However, it might be possible to assign points for continuous/ordered factor.

Advantages of Using Fitted Model

- Can *predict* at any point in region of interest
- Can *optimize* for best setting of factors
- Can *graphically represent* relation between factors and response

Box and Draper: Important Properties of Design

1. Generate information throughout region of interest
2. Ensure that fitted value is as close as possible to true value (or next value).
3. Make it possible to detect lack of fit
4. Performed in blocks
5. Designs of increasing order in sequence

   Linear *then* quadratic *then* cubic....

6. Robust and efficient (compared to what?)
7. Simple patterns for visual inspection (model)
8. Simple to calculate (fitting)
9. Orthogonality: designs have parameter estimators which are uncorrelated
10. Rotatability: variance of prediction depends only on distance from center of design.

Last two can be hard to achieve

Optimal Designs

Need more than "intuitive" understanding of which design is best

1. Most "established" designs have good properties.
2. Need basis for suggesting designs in "non-standard" situations.

Simple Example: Straight Line Through Origin

Set-up

- n observations of Response Y
- $\mathrm{E}(Y) = \beta x$ for $x \in [-1, 1]$ (design region)
- Assume observations uncorrelated with constant variance σ^2

Design

- Set of values: $\{x_1, \ldots, x_n\}$
- Distribution $M(\eta)$: $\{w_1, \ldots, w_n\}$ which show proportion of points at each value $x_1, \ldots, x_n$, respectively.

- Second Design Moment: $M(\eta) = \sum w_i x_i^2$

$$\text{Var}(\hat{\beta}) = \sigma^2/n\{M(\eta)\}^{-1}$$

Typically, $\{w_1, \ldots, w_n\} = \{1/n, \ldots, 1/n\}$.

Often interested in variance of expected value of Y at $x_0 \in [-1, 1]$.

$$\text{Var}(\hat{Y}_0) = \frac{\sigma^2}{n}\{M(\eta)\}^{-1}x_0^2 \equiv \frac{\sigma^2}{n}d(x_0, \eta)$$

Two types of optimality:

$$\text{minimize Var}(\hat{\beta}) \text{ (same as } \max M(\eta))$$
$$\text{or}$$
$$\text{minimize } \max_{x_0} \text{Var}(\hat{Y}_0)$$

Solution: Put all mass at ±1 ($\min_\eta \max_{x_0} d(x_0, \eta) = 1$)

Straight Line with Intercept

$$\text{E}(Y) = \beta_0 + \beta_1 x$$
$$M(\eta) = \begin{pmatrix} 1 & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix}$$

We have that

$$\text{Cov}(\hat{\beta}) = \{M(\eta)\}^{-1}\sigma^2/n$$

Result for even values of n:

> $det(M)$ and $\text{Var}(\hat{\beta}_1)$ minimized with design puts $n/2$ points at 1 and $n/2$ points at -1 (or as far apart as possible)

Variance of estimated response at point x_0:

$$d(x_0, \eta)\sigma^2/n,$$

where

$$d(x, \eta) = (1\, x)\{M(\eta)\}^{-1}(1\, x)^T$$

For symmetrical, 2-point design

$$\min \max_x d(x, \eta) = 2$$

- Any design with $\bar{x} = 0$ minimizes $\text{Var}(\hat{\beta}_0)$
- Design with equal mass at ±1 minimizes $\text{Var}(\hat{\beta}_1)$ and

$$\max_{x_0} \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$$

Critique

- Need precise specification of model
- Optimal design does not check linearity of regression function.

Check

Put some points at 0.
Say

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

- Optimal design for estimating β_1: no points at 0
- Optimal design for estimating β_2: half points at 0
- Optimal design for minimizing $det(\mathrm{Cov}(\hat{\beta}))$ and $\max_{x_0}(d(x_0, \eta))$: 1/3 points at 0

Need to balance relative importance of estimating β_1 and testing adequacy of linear model.

Alternative: Space-filling Designs

Suppose design points spreading uniformly over $[-1, 1]$

> Get 3-fold increase in variance of estimated slope

Trade-off between having model and exploring space

General Theory

Have a design region and linear model

$$\mathrm{E}(Y) = \beta^T f(x)$$

for all x in design region

- β – $p \times 1$ vector of unknown parameters
- $f(x)$ – $p \times 1$ vector of functions of x

  Example: $\{1, x, x^2, \ldots, x^{p-1}\}$
- η – $n \times 1$ vector of weights $(w_1, \ldots, w_n)$
- n observations with constant variance σ^2
- Use LS estimates of β

Moment matrix

$$M(\eta) = \sum w_i f(x_i) f(x_i)^T$$
$$\text{Cov}(\hat{\beta}_{LS}) = \frac{\sigma^2}{n} M(\eta)$$

Variance of estimated response at point x is

$$\frac{\sigma^2}{n} d(x, \eta) = \frac{\sigma^2}{n} f(x)^T \left\{ M(\eta) \right\}^{-1} f(x)$$

Optimality Criteria

A design is D-optimal / G-optimal if the distribution maximizes $det(M(\eta))$ / min max$_x$ variance of estimator

General Equivalence Theorem

If design η^* is D-optimal, then it is also G-optimal with

$$\max_x d(x, \eta^*) = p,$$

the number of parameters.

Other Results

If design is optimal on p points, then all points have equal mass $1/p$ (i.e. if you have $2p$ EU's, each point gets 2 experimental units).

Other Optimality Criteria

- A-optimality – minimize $trace(M^{-1})$
- E-optimality – minimize variance of least well-determined set of contrasts

Applying Optimality Criterion for General Model/Design Region

Two Situations:

1. Some observations already available and need to supplement the design
2. Only region and model are available.

Usual Idea

- Have an initial design (typically, a plausible design with many points)
- Compute $d(x, \eta_0)$ for initial design.
- Add design mass where d is large.
- Iterate until optimality criterion reached.
- Put in n points with approximately the same distribution as optimal.

Montgomery (page 439) has details for such algorithms

Extensions

- Iterative optimization (Response Surface Models)
- Integration
- Mixture Experiments (points are restricted to smaller-dimensional space)
- Experiments with both qualitative and quantitative factors
- Restricted Design regions
- Non-linear models
- Incorporating a prior distribution
- Discrimination between models

Regression Approach to ANOVA

- ANOVA model:

$$y_{i,j} = \mu + \tau_i + \epsilon_{i,j} \left\{ \begin{array}{l} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n_i \end{array} \right.$$

- Can compare it to regression model

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

  where

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

  and, for example,

$$x_{i,j} = \left\{ \begin{array}{cl} 1 & \text{if } j\text{th observation is from treatment } i \\ 0 \text{ (or } -1) & \text{otherwise} \end{array} \right.$$

- Use β's for regression, μ and τ's for ANOVA.

Fitting Regression Models

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

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

General Regression Significance Test

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

  For example, $\beta_A = \beta_0$, and $\beta_B = (\beta_1, \ldots, \beta_a)$

- Want to test $\text{H}_0 : \beta_B = 0$
  - Define $\beta^* = (\beta_A, 0)'$.
  - Fit $y = X\beta$ (Full) and $y = X\beta^*$ (Reduced)

$$F_0 = \frac{(SS_E(Reduced) - SS_E(Full))/(df_R - df_F)}{SS_E(Full)/df_F}$$

  - If $\epsilon \sim N(0, \sigma^2)$, and H_0 true, $F_0 \sim F_{df_R - df_F, df_F}$.
- Can generate F-test for any subset of β.

Orthogonality

- Recall orthogonal contrasts
  - If C and D contrasts, $\sum c_i d_i = 0$.
  - If a treatments, can construct $a-1$ orthogonal contrasts
  - SS_{Trt} separated in $a-1$ independent SS with 1 degree of freedom
  - Order of computation does not affect results.
- Orthogonal regression variables
  - Regressors x_1 and x_2 are orthogonal if $\sum x_{1,j}x_{2,j} = 0$ ($X'X = I$).
  - Order of partitioning does not affect SS results.
  - Blocks of variables can be orthogonal
  - "Orthogonal" experimental designs form sets of orthogonal regressors

    Effect size doesn't depend on presence of other factor
    Each combination appears same number of times.

Application

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

Trt 1	Trt 2	Trt 3
20	24	26
23	26	27
22	25	27

$\bar{y}_{..} = 24.44$, $\bar{y}_{1.} = 21.67$, $\bar{y}_{2.} = 25.00$, $\bar{y}_{3.} = 26.67$

- SAS converts `glm` commands into matrices.
- Includes column for intercept and **ALL** treatments.
- Defining $x_{i,j}$ as before

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

- Over-parameterized model ($rank(X'X) = a$).

- Uses generalized inverse (similar to setting $\beta_3 = 0$)
- Setting $\beta_3 = 0$ amounts to removing last column of X.
- When observation from

$$\begin{array}{lcl} \text{Trt 1} & \rightarrow & E(y_j) = \beta_0 + \beta_1 \\ \text{Trt 2} & \rightarrow & E(y_j) = \beta_0 + \beta_2 \\ \text{Trt 3} & \rightarrow & E(y_j) = \beta_0 \end{array}$$

- Can show the linear regression estimates are

$$\begin{array}{rcl} \hat{\beta}_0 & = & \bar{y}_{3.} \\ \hat{\beta}_1 & = & \bar{y}_{1.} - \bar{y}_{3.} \\ \hat{\beta}_2 & = & \bar{y}_{2.} - \bar{y}_{3.} \end{array}$$

- In terms of ANOVA model parameters

$$\begin{array}{rcl} \beta_0 & = & \mu + \tau_3 \\ \beta_1 & = & \tau_1 - \tau_3 \\ \beta_2 & = & \tau_2 - \tau_3 \end{array}$$

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

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

- When observation from

$$\begin{array}{lcl} \text{Trt 1} & \rightarrow & E(y_j) = \beta_0 + \beta_1 \\ \text{Trt 2} & \rightarrow & E(y_j) = \beta_0 + \beta_2 \\ \text{Trt 3} & \rightarrow & E(y_j) = \beta_0 - \beta_1 - \beta_2 \end{array}$$

- The parameter estimates are

$$\begin{array}{rcl} \hat{\beta}_0 = \bar{y}_{..} & \rightarrow & \beta_0 = \mu \\ \hat{\beta}_1 = \bar{y}_{1.} - \bar{y}_{..} & \rightarrow & \beta_1 = \tau_1 \\ \hat{\beta}_2 = \bar{y}_{2.} - \bar{y}_{..} & \rightarrow & \beta_2 = \tau_2 \end{array}$$

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;
```

```
The GLM 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
Source                     DF          Squares    Mean Square   F Value   Pr > F
Model                       2     38.88888889    19.44444444     15.91   0.0040
Error                       6      7.33333333     1.22222222
Corrected Total             8     46.22222222

R-Square     Coeff Var      Root MSE    response Mean
0.841346      4.522670      1.105542         24.44444

                                            Standard
Parameter           Estimate                   Error    t Value    Pr > |t|
Intercept        26.66666667 B      0.63828474      41.78      <.0001
trt       1      -5.00000000 B      0.90267093      -5.54      0.0015
trt       2      -1.66666667 B      0.90267093      -1.85      0.1144
trt       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.  Terms whose
      estimates are followed by the letter 'B' are not uniquely
      estimable.


------------------------------Model 1-----------------------------
                       Analysis of Variance
                                 Sum of         Mean
Source                 DF       Squares       Square  F Value  Pr > F
Model                   2      38.88889     19.44444    15.91  0.0040
Error                   6       7.33333      1.22222
Corrected Total         8      46.22222

Root MSE             1.10554    R-Square     0.8413
Dependent Mean      24.44444    Adj R-Sq     0.7885
Coeff Var            4.52267

                       Parameter Estimates
                    Parameter      Standard
Variable     DF       Estimate          Error    t Value     Pr > |t|
Intercept     1       26.66667        0.63828      41.78       <.0001
x1            1       -5.00000        0.90267      -5.54       0.0015
x2            1       -1.66667        0.90267      -1.85       0.1144


------------------------------Model 2-----------------------------
              Analysis of Variance
                                 Sum of         Mean
Source                 DF       Squares       Square  F Value  Pr > F
Model                   2      38.88889     19.44444    15.91  0.0040
Error                   6       7.33333      1.22222
Corrected Total         8      46.22222

Root MSE             1.10554    R-Square     0.8413
Dependent Mean      24.44444    Adj R-Sq     0.7885
```

```
Coeff Var            4.52267

                         Parameter Estimates
                      Parameter       Standard
Variable     DF        Estimate          Error    t Value    Pr > |t|
Intercept     1        24.44444        0.36851      66.33      <.0001
x1            1        -2.77778        0.52116      -5.33      0.0018
x2            1         0.55556        0.52116       1.07      0.3274
```

Summary of Two Models

- In both cases $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$, similar to usual H_0.
- Model # 1: Similar to SAS

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

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

- Model # 2

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

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

Estimable Functions

- For some combinations of β's, same answer across models.

- Here are several examples, based on previous models:

$$\begin{aligned}\beta_0+\beta_1 &= \mu+\tau_1\\ \beta_0+\beta_2 &= \mu+\tau_2\\ \beta_1-\beta_2 &= \tau_1-\tau_2\\ \beta_1-\beta_3 &= \tau_1-\tau_3\end{aligned}$$

- These unique combinations known as *estimable functions*.
  - Can uniquely estimate treatment differences
  - Can uniquely estimate treatment means (cell model)
  - Cannot uniquely estimate treatment effects, grand mean (without constraints)
  - Linear combinations of the X matrix are estimable.

$$\begin{aligned}\text{Like Model 1: } & \mu=6,\tau_1=3,\tau_2=-6,\tau_3=0\\ \text{Like Model 2: } & \mu=5,\tau_1=4,\tau_2=-5,\tau_3=1\end{aligned}$$

```
proc glm data = one;  /* How to get grand/treatment means */
 class trt;
 model response = trt/e;

The GLM Procedure

General Form of Estimable Functions

Effect          Coefficients
Intercept       L1
trt       1     L2
trt       2     L3
trt       3     L1-L2-L3
```

Unbalanced/Non-orthogonal Designs

Advantage of Balance:

- Easier to analyze
- Equal amounts of information about conditions to compare
- Assumptions matter less (unless you're lucky)
- No distinction between hypothesis tests and prediction

Reasons unbalanced designs may arise

- Blocking factors: can't recruit balanced design
- Treatment factors: budget, missing data
- Important question: missingness mechanism independent of treatments? (if yes, then things get more complicated)

Unweighted vs Marginal Means:

- **Unweighted mean:**
  - each cell means is weighted by number in each cell
  - grand mean – mean all observations
  - Used in Type III SS calculation
- **Weighted mean:**
  - Each cell mean is unweighted
  - grand mean – mean of all cell means
  - Used in Type I SS calculation

First Approach: Regression

```
Call: lm(formula = error ~ strain + env + env * strain)

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  80.6111   2.0137    40.0321   0.0000
    strain1  -8.1875   2.5421    -3.2207   0.0067
    strain2  -8.8681   1.3787    -6.4324   0.0000
        env   4.7500   2.0137     2.3589   0.0346
 strain1env  -2.1458   2.5421    -0.8441   0.4139
 strain2env   2.5625   1.3787     1.8587   0.0859

Multiple R-Squared: 0.8041
----------------------------------------------------------
Call: lm(formula = error ~ env + strain + env * strain)
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  80.6111   2.0137    40.0321   0.0000
        env   4.7500   2.0137     2.3589   0.0346
    strain1  -8.1875   2.5421    -3.2207   0.0067
    strain2  -8.8681   1.3787    -6.4324   0.0000
 envstrain1  -2.1458   2.5421    -0.8441   0.4139
 envstrain2   2.5625   1.3787     1.8587   0.0859

Multiple R-Squared: 0.8041
```

```
---------------------------------------------------------
Call: lm(formula = error ~ strain * env + env + strain)

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  80.6111   2.0137    40.0321   0.0000
    strain1  -8.1875   2.5421    -3.2207   0.0067
    strain2  -8.8681   1.3787    -6.4324   0.0000
        env   4.7500   2.0137     2.3589   0.0346
 strain1env  -2.1458   2.5421    -0.8441   0.4139
 strain2env   2.5625   1.3787     1.8587   0.0859

Multiple R-Squared: 0.8041


Analysis of Variance Table

Response: error
Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value     Pr(F)
strain      2  3136.132 1568.066 21.48413 0.0000757
env         1   425.911  425.911  5.83542 0.0311575
env:strain  2   332.913  166.456  2.28062 0.1415864
Residuals  13   948.833   72.987

           Df Sum of Sq  Mean Sq  F Value     Pr(F)
env         1   154.501  154.501  2.11682 0.1694074
strain      2  3407.543 1703.771 23.34343 0.0000498
env:strain  2   332.913  166.456  2.28062 0.1415864
Residuals  13   948.833   72.987

           Df Sum of Sq  Mean Sq  F Value     Pr(F)
strain      2  3136.132 1568.066 21.48413 0.0000757
env         1   425.911  425.911  5.83542 0.0311575
strain:env  2   332.913  166.456  2.28062 0.1415864
Residuals  13   948.833   72.987
```

Regression and Types of Sums of Squares

- Testing if regression coefficient belongs in model – doesn't depend on order in model
- Estimates of effects don't depend on order
- Useful in building model for prediction.
- However, Type I SS does depend on order of terms in model

Conditional Sums of Squares

$SS(B|1, A)$ means SS of B adjusted for grand mean and A.
MS in adjusted estimator estimates $\sigma^2_{B|A}$.

Conditioning most useful for inference:

$$r\frac{(SS_E(\text{reduced}) - SS_E(\text{full}))}{SS_E(\text{full})} \sim F$$

Different types of SS based on different conditioning

Effect	Type I (sequential)	Type II (partial)	Type III/Type IV
env	$R(env\|\mu)$	$R(env\|\mu, trait)$	$R(env\|\mu, trait, env * trait)$
trait	$R(trait\|\mu)$	$R(trait\|\mu, env)$	$R(trait\|\mu, env, env * trait)$
env×trait	$R(env * trait\|\mu, env, trait)$	$R(env * trait\|\mu, env, trait)$	$R(env * trait\|\mu, env, trait)$

- Hypothesis test for $R(env|\mu)$: Does including the `env` factor make the model fit significantly better than just estimating with the grand mean?
- Hypothesis test for $R(env|\mu, trait, env * trait)$: Does including the `env` factor make the model fit significantly better than the model with `trait`, and the `env*trait` interaction?
- Notice that only in `Type I` does the SS add to `Model SS`.
- Results in `Type II` give results of getting rid of higher-order terms first.
- `Type IV` is considered more "valid" when there are empty cells.

Which one is best?

"It depends"

- If cell sizes represent population sizes, use Type I (keep factor of interest first and ignore other factors).
- Is there no interaction? Type II is more powerful.
- Usually, use Type III.