STAT 525

Chapter 16
Single-Factor Studies

Dr. Qifan Song

One-Way ANOVA

- Response variable Y is again continuous
- Explanatory variable is *categorical*
  - Often called a factor
  - The possible values are its levels
- A generalization of the independent two-sample t-test (i.e., can be used when there are more than two levels)

ANOVA vs. Regression

- One-way ANOVA a special case of regression using indicator variables

- Recall in comparing regression lines, indicator variables were used to describe differences in intercepts (i.e, means)

- Consider the linear model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ involving three groups where X_1 is the indicator for group 1 and X_2 is the indicator for group 2
  - Group 1 : $Y_i = \beta_0 + \beta_1 + \varepsilon_i = \mu_1 + \varepsilon_i$
  - Group 2 : $Y_i = \beta_0 + \beta_2 + \varepsilon_i = \mu_2 + \varepsilon_i$
  - Group 3 : $Y_i = \beta_0 + \varepsilon_i = \mu_3 + \varepsilon_i$

- Indicators remove “linear” structure among means

The Data / Notation

- Y is the response variable
- X is the factor with r levels. These levels are often called groups or treatments.
- Let Y_{ij} be the
  - j^{th} observation ($j = 1, 2, ..., n_i$)
  - in the i^{th} group ($i = 1, 2, ..., r$)

Example (Page 685)

- Kenton Food Company wants to test four different package designs for a new breakfast cereal
- Twenty "similar" stores were selected to be part of the experiment
- Package designs randomly and equally assigned to stores. Fire hit one store so it was dropped
- Y is the number of cases sold
- X is the package design with $r = 4$ levels
  - $i = 1, 2, 3, 4$
  - $j = 1, 2, .., n_i$ where $n_i = 5, 5, 4, 5$ respectively
  - will use n when n_i constant

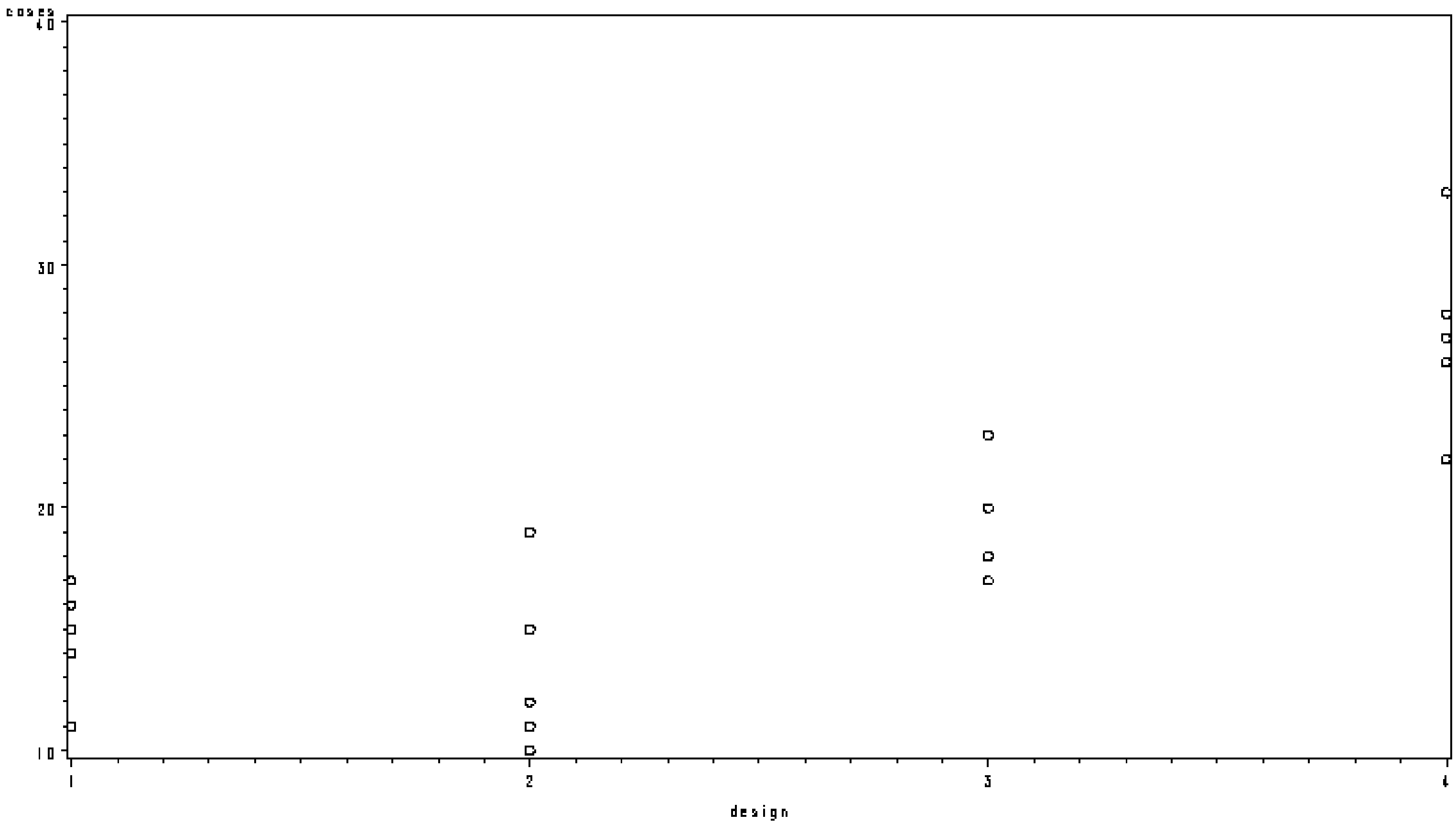
The Data

```
data a1;
    infile 'u:\.www\datasets525\CH16TA01.TXT';
    input cases design store;
proc print; run; quit;
```

Obs	cases	design	store
1	11	1	1
2	17	1	2
3	16	1	3
4	14	1	4
5	15	1	5
6	12	2	1
7	10	2	2
8	15	2	3
9	19	2	4
10	11	2	5
11	23	3	1
12	20	3	2
13	18	3	3
14	17	3	4
15	27	4	1
16	33	4	2
17	22	4	3
18	26	4	4
19	28	4	5

Scatterplot

```
symbol1 v=circle i=none;
proc gplot data=a1;
    plot cases*design/frame;
run; quit;
```



X-axis has no numerical meaning.

The Model

- Same assumptions as regression except for the linear relationship between X and Y
- All observations are assumed independent
- All observations are normally distributed with
  - means which may depend on the levels of the factors
  - constant variance
- Often presented in terms of cell means or factor effects

The Cell Means Model

- Expressed numerically

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

  where μ_i is the theoretical mean of all observations at level i (or in cell i)

- The ε_{ij} are iid $N(0, \sigma^2)$ which implies the Y_{ij} are independent $N(\mu_i, \sigma^2)$

- Parameters

  - $\mu_1, \mu_2, ..., \mu_r$

  - σ^2

Primary Question

- Does the explanatory variable X help explain Y?
- Since the factor levels only affect the cell means we can similarly ask ...
- Does μ_i depend on i?
  - $H_0 : \mu_1 = \mu_2 = ... = \mu_r = \mu$
  - H_a : at least one μ_i different

Estimates / Inference

- Derive the following result via matrix form of linear regression
- Estimate μ_i by the sample mean of the observations at level i

$$\widehat{\mu}_i = \overline{Y}_{i.}$$

- For each level i, also estimate of the variance

$$s_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2/(n_i - 1)$$

- These s_i^2 are combined to estimate σ^2
  - If n_i were constant, could compute s^2 by averaging the s_i^2's
  - More general formula pools s_i^2 using weights proportional to sample size (i.e., df)

$$s^2 = \frac{\sum_{i=1}^{r} (n_i - 1)s_i^2}{\sum_{i=1}^{r} (n_i - 1)} = \frac{\sum_{i=1}^{r} (n_i - 1)s_i^2}{n_T - r}$$

    where n_T is the total number of obs

ANOVA Table

- Similar ANOVA table construction
- Plug in $\overline{Y}_{i.}$ as fitted value

Source of Variation	df	SS
Model	$r-1$	$\sum_{i=1}^{r} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$
Error	$n_T - r$	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2$
Total	$n_T - 1$	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$

- Note that

$$\overline{Y}_{..} = \sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij}/n_T \qquad \overline{Y}_{i.} = \sum_{j=1}^{n_i} Y_{ij}/n_i$$

- $SSM = SS(B)$, aka the between-group variation;
- $SSE = SS(W)$, aka the within-group variation.

Expected Mean Squares (EMS)

- All means squares are random variables
- Can show $E(\text{MSE}) = \sigma^2$ (page 696)
- Can also show (page 697)

$$E(\text{MSR}) \;=\; \sigma^2 + \frac{\sum n_i(\mu_i - \mu_.)^2}{r-1}$$

where $\mu_. = \frac{\sum n_i \mu_i}{n_T}$

- If H_0 true, MSR unbiased estimate of σ^2. More specifically, SSE/σ^2 and SSR/σ^2 are independent χ^2 distribution
- In more complicated ANOVA models, EMS (Hasse diagram; STAT 514) guides us how to construct F tests

Example (Page 685) – Use PROC GLM in SAS

```
/* GLM: Uses least squares method to fit general linear models, and     */
/*      provides regression, ANOVA, ANCOVA, MANCOVA, partial correlation */
/* Automatically create indicator variable by class statement           */
proc glm data=a1;
    class design;
    model cases=design;
    means design;
    lsmeans design / stderr;
run; quit;
```

```
                               Sum of
Source             DF         Squares  Mean Square  F Value  Pr > F
Model               3     588.2210526  196.0736842    18.59  <.0001
Error              15     158.2000000   10.5466667
Corrected Total    18     746.4210526

R-Square     Coeff Var      Root MSE    cases Mean
0.788055      17.43042      3.247563      18.63158

Source             DF       Type I SS  Mean Square  F Value  Pr > F
design              3     588.2210526  196.0736842    18.59  <.0001

Source             DF     Type III SS  Mean Square  F Value  Pr > F
design              3     588.2210526  196.0736842    18.59  <.0001
```

The GLM Procedure

Level of design	N	cases Mean	cases Std Dev
1	5	14.6000000	2.30217289
2	5	13.4000000	3.64691651
3	4	19.5000000	2.64575131
4	5	27.2000000	3.96232255

Least Squares Means

design	cases LSMEAN	Standard Error	Pr > |t|
1	14.6000000	1.4523544	<.0001
2	13.4000000	1.4523544	<.0001
3	19.5000000	1.6237816	<.0001
4	27.2000000	1.4523544	<.0001

plus some plots.

- MEANS only uses the observations from a specific group
  - $4 \times 2.30^2 + 4 \times 3.65^2 + 3 \times 2.65^2 + 4 \times 3.96^2 = 158.24$. Except for rounding, this is equal to SSE.
  - $19 - 4 = 15$, which is the df error in the ANOVA table.
- LSMEANS uses all the observations and least squares method
  - $SE_i = \sqrt{MSE/n_i}$.

Example (Page 685) – Use PROC MIXED in SAS

```
/* MIXED: generalizes the linear models in PROC GLM & fits linear mixed models */
proc mixed data=a1;
    class design;
    model cases=design;
    lsmeans design;
run; quit;
```

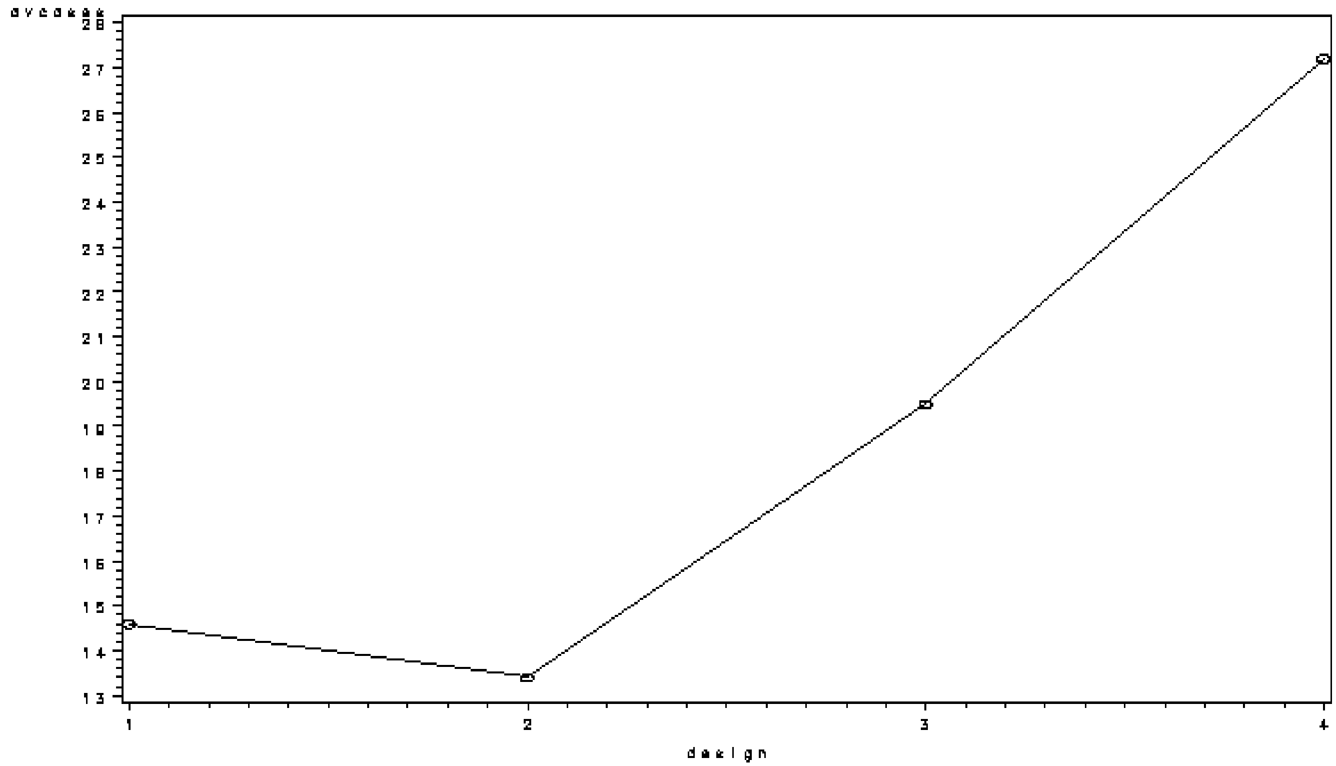
```
        Type 3 Tests of Fixed Effects
              Num     Den
Effect         DF      DF    F Value    Pr > F
design          3      15      18.59    <.0001

                Least Squares Means
                      Standard
design    Estimate       Error      DF    t Value    Pr > |t|
1          14.6000      1.4524      15      10.05      <.0001
2          13.4000      1.4524      15       9.23      <.0001
3          19.5000      1.6238      15      12.01      <.0001
4          27.2000      1.4524      15      18.73      <.0001
```

Scatterplot of Means

Generated by `lsmeans design/plot=meanplot(join);` in glm procedure, or manually:

```
proc means data=a1;
    var cases; by design;
    output out=a2 mean=avcases;
symbol1 v=circle i=join;
proc gplot data=a2;
    plot avcases*design/frame;
run; quit;
```



The Factor Effects Model

- A reparameterization of the cell means model
- A very useful way of looking at more complicated ANOVA models (i.e., more than one factor)
- Null hypotheses are easier to state
- Expressed numerically

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

- Parameters
  - $\tau_1, \tau_2, ..., \tau_r$
  - μ, σ^2
- Factor effects model has $r+2$ parameters while the cell means model has $r+1$ parameters

Model Identifiability

- Consider $r = 3$ with $\mu_1 = 10, \mu_2 = 0$, and $\mu_3 = 20$
  - $\mu = 0, \tau_1 = 10, \tau_2 = 0, \tau_3 = 20$
  - $\mu = 10, \tau_1 = 0, \tau_2 = -10, \tau_3 = 10$
  - $\mu = 100, \tau_1 = -90, \tau_2 = -100, \tau_3 = -80$
- Factor effects model has non-unique solution
- Solution: put constraints on τ_i's to reduce the parameters number by 1
- Examples of constraints
  - $\tau_r = 0$ (SAS approach)
  - $\sum \tau_i = 0$ (conceptual approach)
- Constraints get a bit more complicated when n_i not constant (pages 709-710) but with same concept

Consequences of Constraints

- Consider $r = 3$ with $n_i = n$

- Factor effects model with $\sum \tau_i = 0$

$$E(\overline{Y}_{..}) = \frac{3\mu + \sum \tau_i}{3} = \mu$$
$$E(\overline{Y}_{i.}) = \mu + \tau_i$$

  – μ is the grand mean

  – τ_i is the effect of the i^{th} factor

- Factor effects model with $\tau_r = 0$

$$E(\overline{Y}_{3.}) = \mu$$
$$E(\overline{Y}_{1.} - \overline{Y}_{3.}) = \mu + \tau_1 - \mu = \tau_1$$

  – μ is the mean of the r^{th} group

  – τ_i is the difference between the means of group i and group r

- Different constraints result in different parameter / parameter estimates

- Many estimates, however, are the same regardless of constraint

  - $\widehat{\mu} + \widehat{\tau}_1 =$ trt 1 mean

  - $\widehat{\mu} + \widehat{\tau}_3 =$ trt 3 mean

  - $\widehat{\tau}_1 - \widehat{\tau}_3 =$ difference in trt 1 and trt 3

  - $\widehat{\tau}_1 - \widehat{\tau}_2 =$ difference in trt 1 and trt 2

- These are primarily the ones of interest

Hypotheses

$H_0 : \mu_1 = \mu_2 = ... = \mu_r = \mu$

H_a : at least one μ_i different

is translated into

$H_0 : \tau_1 = \tau_2 = ... = \tau_r = 0$

H_a : at least one $\tau_i \neq 0$

Regression Approach

- We can use multiple regression to produce results based on the factor effects model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

- Consider the restriction $\sum \tau_i = 0$
- Because of this restriction, effectively there are $r-1$ regression coefficients /parameters

$\sum \tau_i = 0 \rightarrow \tau_r = -\tau_1 - \tau_2 - ... - \tau_{r-1}$

- Define k-th indicator variable ($k = 1, 2, \cdots, r-1$)

$$X_{ijk} = \begin{cases} 1, & \text{factor level at } k\text{, i.e., } i = k \\ -1, & \text{factor level at } r\text{, i.e., } i = r \\ 0, & \text{otherwise} \end{cases}$$

- Multiple regression model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + ... + \beta_{r-1} X_{ij,r-1} + \varepsilon_{ij}$$

  – For level i $(1 \le i \le r-1)$

$$Y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij}$$

  – For level r

$$Y_{ij} \;=\; \beta_0 - \beta_1 - \beta_2 - ... - \beta_{r-1} + \varepsilon_{ij}$$

- Perfectly match $\mu = \beta_0$ and $\tau_i = \beta_i$ $(1 \le i \le r-1)$

- Solve all β_i via the multiple linear regression approach.

- $\widehat{\mu} = b_0 = \sum_{i=1}^{r} \overline{Y}_{i.}/r$ (if n_i's are not constant, then $b_0 \neq \overline{Y}_{..}$), $\hat{\tau}_i = b_i = \overline{Y}_{i.} - b_0$.

```
/* Code Indicator Variables */
data a1; set a1;
   x1=(design eq 1)-(design eq 4);
   x2=(design eq 2)-(design eq 4);
   x3=(design eq 3)-(design eq 4);
proc print data=a1; run; quit;
```

Obs	cases	design	store	x1	x2	x3
1	11	1	1	1	0	0
2	17	1	2	1	0	0
3	16	1	3	1	0	0
4	14	1	4	1	0	0
5	15	1	5	1	0	0
6	12	2	1	0	1	0
7	10	2	2	0	1	0
8	15	2	3	0	1	0
9	19	2	4	0	1	0
10	11	2	5	0	1	0
11	23	3	1	0	0	1
12	20	3	2	0	0	1
13	18	3	3	0	0	1
14	17	3	4	0	0	1
15	27	4	1	-1	-1	-1
16	33	4	2	-1	-1	-1
17	22	4	3	-1	-1	-1
18	26	4	4	-1	-1	-1
19	28	4	5	-1	-1	-1

```
proc reg data=a1;
   model cases=x1 x2 x3;
run; quit;
```

```
                 Analysis of Variance
                         Sum of        Mean
Source              DF    Squares      Square  F Value  Pr > F
Model                3  588.22105   196.07368    18.59  <.0001
Error               15  158.20000    10.54667
Corrected Total     18  746.42105

Root MSE                3.24756    R-Square     0.7881
Dependent Mean         18.63158    Adj R-Sq     0.7457
Coeff Var              17.43042

                        Parameter Estimates
                    Parameter   Standard
Variable     DF    Estimate       Error    t Value    Pr > |t|
Intercept     1    18.67500     0.74853      24.95      <.0001
x1            1    -4.07500     1.27081      -3.21      0.0059
x2            1    -5.27500     1.27081      -4.15      0.0009
x3            1     0.82500     1.37063       0.60      0.5562
```

- The mean of the means is 18.675
- The treatment means are $18.675 - 4.075 = 14.6$, $18.675 - 5.275 = 13.4$, $18.675 + 0.825 = 19.5$, and $18.675 + 4.075 + 5.275 - 0.825 = 27.2$
- The same output as from `PROC GLM` before

SAS Regression Approach

- `class` statement constructs the following r indicator variables

$$X_{ijk} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$$

- Because of the intercept (column of 1's) there is complete dependence ($\mathbf{X}'\mathbf{X}$ doesn't have an inverse)

$$\mathbf{1} = \mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_r$$

- SAS computes *generalized inverse* in its place. (Generalized inverse of A, A^-, satisfies $AA^-A = A$)

- $b = (\mathbf{X}'\mathbf{X})^-(\mathbf{XY})$ satisfies $(\mathbf{X}'\mathbf{X})b = (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^-(\mathbf{XY}) = (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^-(\mathbf{X}'\mathbf{X})\tilde{\beta} = (\mathbf{X}'\mathbf{X})\tilde{\beta} = \mathbf{XY}$, for some $\tilde{\beta}$.

- Generalized inverse is not unique, and SAS choose the a particular one such that $\widehat{\tau}_r = 0$.

Example (Page 685)

```
proc glm data=a1;
    class design;
    model cases=design / xpx inverse solution;
run; quit;
```

The X'X Matrix

	Int	d1	d2	d3	d4	cases
Int	19	5	5	4	5	354
d1	5	5	0	0	0	73
d2	5	0	5	0	0	67
d3	4	0	0	4	0	78
d4	5	0	0	0	5	136
cases	354	73	67	78	136	7342

X'X Generalized Inverse (g2)

	Int	d1	d2	d3	d4	cases
Int	0.2	-0.2	-0.2	-0.2	0	27.2
d1	-0.2	0.4	0.2	0.2	0	-12.6
d2	-0.2	0.2	0.4	0.2	0	-13.8
d3	-0.2	0.2	0.2	0.45	0	-7.7
d4	0	0	0	0	0	0
cases	27.2	-12.6	-13.8	-7.7	0	158.2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	588.2210526	196.0736842	18.59	<.0001
Error	15	158.2000000	10.5466667		
Corrected Total	18	746.4210526			

R-Square	Coeff Var	Root MSE	cases Mean
0.788055	17.43042	3.247563	18.63158

Source	DF	Type I SS	Mean Square	F Value	Pr > F
design	3	588.2210526	196.0736842	18.59	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
design	3	588.2210526	196.0736842	18.59	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		27.20000000	B	1.45235441	18.73	<.0001
design	1	-12.60000000	B	2.05393930	-6.13	<.0001
design	2	-13.80000000	B	2.05393930	-6.72	<.0001
design	3	-7.70000000	B	2.17853162	-3.53	0.0030
design	4	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.

Interpretation

- Generalized Inverse Matrix of the form

$$\begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-} & (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-} & \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} \end{bmatrix}$$

- Parameter estimates in upper right corner and SSE in lower right corner

- The intercept in the parameter estimation is actually the mean estimator for the last group.

Chapter Review

- One Way ANOVA
  - Cell means model
  - Factor effects model
- Regression Approach to ANOVA