Topic 17 - Single Factor Analysis of Variance

STAT 525 - Fall 2013

Outline

- One way ANOVA
  - Cell means model
  - Factor effects model

One-way ANOVA

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

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 $(i = 1, 2, ..., r)$

ANOVA vs Regression

- ANOVA a special case of regression using indicator variables
- Recall in comparing regression lines, indicator variables were used to describe differences in intercepts
- 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 : $Y_i = \beta_0 + \varepsilon_i = \mu_3 + \varepsilon_i$
- Allows each level of factor to have different intercept (i.e., mean). There is no linear structure among these means.

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

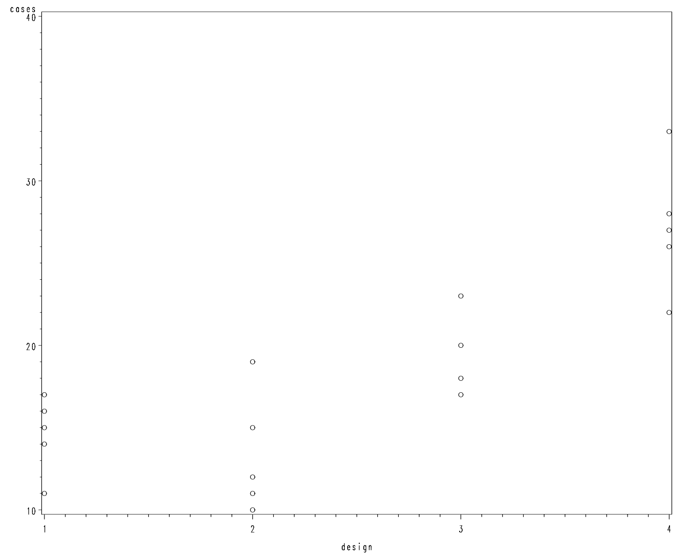
SAS Commands

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

symbol1 v=circle i=none;
proc gplot data=a1;
   plot cases*design;

proc means data=a1;
   var cases; by design;
   output out=a2 mean=avcases;
proc print data=a2;

symbol1 v=circle i=join;
proc gplot data=a2;
   plot avcases*design;
run;
```

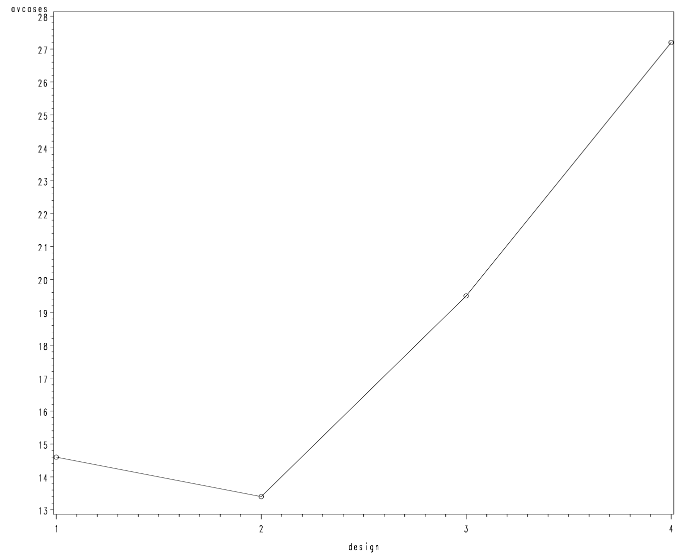

The Data

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



Profile Plot



The Model

- Since a special case of linear regression, same assumptions on errors hold. This implies...
- All observations assumed independent
- All observations Normally distributed with
  - a mean that may depend on the level of the factor
  - constant variance
- Model often presented in terms of the cell means or the 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

- In simple linear regression we ask "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

- Estimate μ_i by the sample mean of the observations at level i

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

- For each level i, also estimate of the variance

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

- These s_i^2 are combined to estimate σ^2
- NOTE: Same summaries computed for independent two-sample t-test

Estimates / Inference

- If n_i were constant, can 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)

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

where n_T is the total number of obs

- NOTE: Do <u>not</u> pool or average s_i's

ANOVA Table

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

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

$$\overline{Y}_{..} = \sum\sum Y_{ij}/n_T$$
$$\overline{Y}_{i.} = \sum Y_{ij}/n_i$$

Expected Mean Squares

- All mean 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
- In more complicated ANOVA models, EMS tell us how to construct F tests

SAS Commands

```
proc glm data=a1;
   class design;
   model cases=design;
   means design;                  ***Don't use***
   lsmeans design / stderr;
run;

proc mixed data=a1;
   class design;
   model cases=design;
   lsmeans design;
run;
```


Output - GLM

```
Class Level Information
Class          Levels    Values
design              4    1 2 3 4

Number of observations    19

                              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
```

Output - GLM

```
The GLM Procedure

Level of          ------------cases------------
design       N          Mean          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
                                 Standard
design    cases LSMEAN              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
```

Note: $4 \times 2.30^2 + 4 \times 3.65^2 + 3 \times 2.65^2 + 4 \times 3.96^2 = 158.24$. Except for rounding, this is equal to SSE. Also, 19-4=15 which is the df error in the ANOVA table.

Output - MIXED

```
                 Model Information
Data Set                      WORK.A1
Dependent Variable            cases
Covariance Structure          Diagonal
Estimation Method             REML         *****
Residual Variance Method      Profile
Fixed Effects SE Method       Model-Based
Degrees of Freedom Method     Residual

             Class Level Information
Class     Levels    Values
design         4    1 2 3 4

            Dimensions
Covariance Parameters              1       *****
Columns in X                       5
Columns in Z                       0
Subjects                           1       *****
Max Obs Per Subject               19       *****
```


Output - MIXED

```
Covariance Parameter Estimates
    Cov Parm     Estimate
    Residual      10.5467

           Fit Statistics
-2 Res Log Likelihood             84.1
AIC (smaller is better)           86.1
BIC (smaller is better)           86.8

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

                Least Squares Means
design    Estimate  Std. 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
```


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

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

The Factor Effects Model

- Parameters
  - $\mu,\ \tau_1, \tau_2, ..., \tau_r$
  - σ^2
- Factor effects model has $r+2$ parameters while the cell means model has $r+1$ parameters
- Overparameterized...not a unique solution
- Consider $r = 3$ with $\mu_1 = 10, \mu_2 = 0$, and $\mu_3 = 20$
  - $\mu = 0, \tau_1 = 10, \tau_2 = 0, \tau_3 = 20$
  - $\mu = 10, \tau_1 = 0, \tau_2 = -10, \tau_3 = 10$
  - $\mu = 100, \tau_1 = -90, \tau_2 = -100, \tau_3 = -80$

The Factor Effects Model

- Because the factor effects model has non-unique solution, we put a constraint on the τ_i's
- Examples of constraint
  - $\tau_r = 0$ (SAS approach)
  - $\sum \tau_i = 0$ (conceptual approach)
- Reduces the number of parameters by 1 so we now have a unique solution

Consequences of Constraint Choice

- Consider $r = 3$ with $n_i = n$
- Factor effects model with constraint $\sum \tau_i = 0$

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

In this case μ is the "grand" mean and τ_i is the effect of the i^{th} factor

- Factor effects model with $\tau_r = 0$

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

In this case μ is the mean of the r^{th} group and τ_i is the difference between the means of group i and group r

Consequences of Constraints

- Different constraints result in different parameter / parameter estimates
- Many estimates, however, are the same regardless of constraint. Recall our example
  - $\hat{\mu} + \hat{\tau}_1$ = trt 1 mean
  - $\hat{\mu} + \hat{\tau}_3$ = trt 3 mean
  - $\hat{\tau}_1 - \hat{\tau}_3$ = difference in trt 1 and trt 3
  - $\hat{\tau}_1 - \hat{\tau}_2$ = difference in trt 1 and trt 2
- These are primarily the ones of interest
- Details a bit more complicated when n_i not constant (pages 709-710) but same concept

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 $\tau_i \neq 0$

Background Reading

- KNNL Section 16.1-16.7
- knnl686.sas
- knnl717.sas
- KNNL Chapter 16.8