Displays for Statistics 5303

Lecture 23

October 28, 2002

Christopher Bingham, Instructor

612-625-7023 (St. Paul)
612-625-1024 (Minneapolis)

Class Web Page
http://www.stat.umn.edu/~kb/classes/5303

Unbalanced data continued

Why is balance important?

The short answer is this.

When data are not balanced,

- Calculation is much harder; you really need a computer program
- The order of terms in the model can make a difference in the SS, at least as computed by MacAnova (type I SS)
- The sums of squares used for testing don't add up to what you might think they should
- You may need one or both of factors A and B, but each can each appear to be insignificant (small F statistics) when they are both in the model.

Severe lack of balance can be considered a form of **multicollinearity**, a problem that arises in multiple regression when predictor variables are highly correlated.

An important advantage of balanced data:

Contrasts going with different terms in the model are **orthogonal**. Examples are for a 2 by 3 design

- Contrasts in different main effects are orthogonal.

	A main effect.		
	B1	B2	B3
A1	1	1	1
A2	-1	-1	-1

B main effect		
B1	B2	B3
-1	0	1
-1	0	1

The sum of products of the 6 values in the left A-main effect contrast times the corresponding 6 values in the right B-main effect contrast is 0.

- Main effect contrasts are orthogonal to interaction contrasts.

	A main effect.		
	B1	B2	B3
A1	1	1	1
A2	-1	-1	-1

AB interact		
B1	B2	B3
-1	0	1
1	0	-1

- Interaction contrasts associated with different interactions terms in the model are orthogonal.

  This can't be illustrated with two factors since there is only one interaction term.

It's really this orthogonality property that results in the order of terms being irrelevant with balanced data, but very important with unbalanced data.

In regression terms orthogonality of different terms is analogous to two predictor variables x_1 and x_2 having zero correlation, that is

$$\sum(x_{1i} - \overline{x}_{1\bullet})(x_{2i} - \overline{x}_{2\bullet}) = 0$$

Example based on Problem 8.1 data from a 5 by 2 factorial experiment..

```
Cmd> data <- read("","pr8.1")
pr8.1     30     3
) A data set from Oehlert (2000) \emph{A First Course in Design
) and Analysis of Experiments}, New York: W. H. Freeman.
)
) Data originally from Hareland, G.~A. and M.~A. Madson (1989).
) ``Barley dormancy and fatty acid composition of lipids
) isolated from freshlyharvested and stored kernels.'' {\em
) Journal of the Institute of Brewing} {\em 95}, 437--442.
)
) Table 8.1, p. 166
) Columns are weeks, water, and response (number of seeds
) germinating).
) Codes 1, 2, ... 5 for weeks are 1, 3, 6, 9, 12 weeks.
) Codes 1, 2 for water are 4, 8 mls.
Read from file "TP1:Stat5303:Data:OeCh08.dat"

Cmd> makecols(data, weeks, water, y)

Cmd> weeks <- factor(weeks); water <- factor(water)

Cmd> tabs(y,weeks,water,count:T) # equal sample sizes
(1,1)          3           3
(2,1)          3           3
(3,1)          3           3
(4,1)          3           3
(5,1)          3           3
```

All n_{ij} are equal ⇒ data are **balanced**.

```
Cmd> anova("y = weeks + water + weeks.water",fstat:T)
Model used is y = weeks + water + weeks.water
              DF          SS          MS           F     P-value
CONSTANT       1      6049.2      6049.2   101.27009   2.845e-09
weeks          4      1321.1      330.28     5.52930   0.0036449
water          1      1178.1      1178.1    19.72321  0.00025098
weeks.water    4      208.87      52.217     0.87416     0.49673
ERROR1        20      1194.7      59.733
```

The interaction is not significant so I will work with the additive model.

Fit a model with **weeks** before **water**:

```
Cmd> anova("y = weeks + water",fstat:T)
Model used is y = weeks + water
              DF          SS          MS           F     P-value
CONSTANT       1      6049.2      6049.2   103.43951  3.5295e-10
weeks          4      1321.1      330.28     5.64775   0.0023801
water          1      1178.1      1178.1    20.14573  0.00015255
ERROR1        24      1403.5      58.481
```

Fit a model with **water** before **weeks**:

```
Cmd> anova("y = water + weeks",fstat:T)
Model used is y = water + weeks
              DF          SS          MS           F     P-value
CONSTANT       1      6049.2      6049.2   103.43951  3.5295e-10
water          1      1178.1      1178.1    20.14573  0.00015255
weeks          4      1321.1      330.28     5.64775   0.0023801
ERROR1        24      1403.5      58.481
```

Lines are in a different order, but SS, MS and F are the same. Also SS_E are the same.

ANOVA is really regression in disguise. As in regression can define SS_{reg} as the sum of squares "explained" by the categorical predictors. SS_{reg} is sometimes called the *model SS*.

The total variation to be explained is $SS_{total} = \sum(y_{ijk} - \overline{y_{\dots}})^2$. This can be viewed as the **residual SS** when you fit the "trivial" model $y_{ijk} = \mu + \varepsilon_{ijk}$

```
Cmd> ss_tot <- sum((y - describe(y,mean:T))^2) # Total SS

Cmd> ss_resid <- SS[4] # Residual SS

Cmd> ss_reg <- ss_tot - ss_resid; ss_reg # Regression SS
(1)      2499.3
```

ss_resid here is the residual SS when you fit the model $y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$ and **ss_reg** is the amount the residual SS was reduced by fitting this model as compared to the trivial model.

The SS computed by MacAnova are *sequential*. Each is the SS "explained" by each term *in addition* to previous terms. So the overall SS explained by **water** and **weeks** is **SS[2] + SS[3]**.

```
Cmd> SS[2] + SS[3]
(1)      2499.3     same as ss_reg
```

This does *not* depend on order, even when data are unbalanced.

Now I modify the data set to make it unbalanced. I copied **y** to **y1** and set **y1[1]** to **MISSING** by **y1[1] <- ?** or **y1[1] <- NA**.

```
Cmd> y1 <- y; y1[1] <- ? # or y1[1] <- NA

Cmd> tabs(y1,weeks,water,count:T)
WARNING: MISSING values in argument 1 to tabs() omitted
(1,1)          2          3
(2,1)          3          3
(3,1)          3          3
(4,1)          3          3
(5,1)          3          3
```

Now $n_{11} = 2$ and all other $n_{ij} = 3$

```
Cmd> ss_tot1 <- sum((y1[-1] - describe(y1[-1],mean:T))^2)

Cmd> ss_tot1 # modified data total SS
(1)      3892.2

Cmd> anova("y1=weeks + water", fstat:T)
Model used is y1=weeks + water
WARNING: cases with missing values deleted
WARNING: summaries are sequential
              DF          SS          MS           F     P-value
CONSTANT       1      5938.8      5938.8    97.64532  9.5671e-10
weeks          4      1333.1      333.27     5.47958   0.0030012
water          1      1160.3      1160.3    19.07713  0.00022514
ERROR1        23      1398.9       60.82

Cmd> ss_resid1 <- SS[4]; ss_resid1# modified data residual SS
      ERROR1
      1398.9

Cmd> ss_reg1 <- ss_tot1 - ss_resid1; ss_reg1
(1)      2493.3
```

It is still the case that SS_{reg} is the sum of the SS for **weeks** and **water**:

```
Cmd> SS[2] + SS[3]
(1)      2493.3
```

Redo the ANOVA with **weeks** after **water**:

```
Cmd> anova("y1=water + weeks",fstat:T)
Model used is y1=water + weeks
WARNING: cases with missing values deleted
WARNING: summaries are sequential
              DF          SS          MS           F      P-value
CONSTANT       1      5938.8      5938.8    97.64532  9.5671e-10
water          1      1263.6      1263.6    20.77537  0.00014007
weeks          4      1229.8      307.45     5.05502   0.0045115
ERROR1        23      1398.9       60.82

Cmd> SS[2] + SS[3]  # same sum = ss_reg
(1)      2493.3
```

Although the SS for **weeks** and **water** still add up to the SS_{reg}, they each differ from the SS in the ANOVA with **water** after **weeks**.

A general principle in regression and ANOVA.

Tests to decide if a quantitative or categorical variable should be in the model are based on how much SS_{resid} is reduced and SS_{reg} is increased when the variable is added to the model after the other terms in the model.

The SS reported by **anova()** for a term is the SS when the associated term is added to the model which includes all the terms that precede it. It is relevant only in a model that has no terms entered after the term under test.

Let's start fitting the trivial model

$y_{ijk} = \mu + \varepsilon_{ijk}$

```
Cmd> anova("y1=1") # y_ijk = mu + e_ijk
Model used is y1=1
WARNING: cases with missing values deleted
                DF         SS          MS
CONSTANT         1     5938.8      5938.8
ERROR1          28     3892.2      139.01

Cmd> rss1 <- SS[2] # save SS_error
```

Note: The error SS is SS_{total}.

Now fit the model $y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$

```
Cmd> anova("y1=water") # y_ijk=mu + alpha_i + e_ijk
Model used is y1=water
WARNING: cases with missing values deleted
WARNING: summaries are sequential
                DF         SS          MS
CONSTANT         1     5938.8      5938.8
water            1     1263.6      1263.6
ERROR1          27     2628.6      97.357

Cmd> rss2 <- SS[3]; rss1 - rss2
      ERROR1
      1263.6      SS_water = difference of Rss's
```

rss1 and **rss2** are the residual SS from the trivial model and the model including **water** but not **weeks**.

Statistics 5303 Lecture 23 October 28, 2002

Now the fit the full additive model with both **water** and **weeks**

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

```
Cmd> anova("y1=water + weeks") # y_ijk=mu+alpha_i+betas_j+e_ijk
Model used is y1=water + weeks
WARNING: cases with missing values deleted
WARNING: summaries are sequential
                DF         SS          MS
CONSTANT         1     5938.8      5938.8
water            1     1263.6      1263.6
weeks            4     1229.8      307.45
ERROR1          23     1398.9       60.82

Cmd> rss3 <- SS[4]; rss2 - rss3
      ERROR1
      1229.8      SS_weeks = difference of Rss's
```

rss3 is the residual SS from this model and SS_{weeks} = **rss2 - rss3** is the reduction in the residual SS by including $\{\beta_j\}$ in the model in addition to μ and $\{\alpha_i\}$.

SS_{weeks} in this ANOVA does fine to compute F to test $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, but SS_{water} is not OK to test H_0: $\alpha_1 = \alpha_2 = 0$

To test H_0: $\alpha_1 = \alpha_2 = 0$, you need to put **water** after **weeks**, so SS_{water} measures how much **water** "explains" that can't be explained by **weeks**:

```
Cmd> anova("y1=weeks + water") # water after weeks
Model used is y1=weeks + water
WARNING: cases with missing values deleted
WARNING: summaries are sequential
              DF          SS          MS
CONSTANT       1      5938.8      5938.8
weeks          4      1333.1      333.27
water          1      1160.3      1160.3
ERROR1        23      1398.9       60.82
```

So the SS to be used in testing are

SS_{water} = 1160.3 from this ANOVA

and SS_{weeks} = 1229.8 from the preceding.

But these no longer add up to the model SS

```
Cmd> ss_reg1
(1)      2493.3      Model SS

Cmd> 1229.8 + 1160.3
(1)      2390.1      Sum of SS used in F-tests
```


Notation

Because there may be several different sums of squares for a factor or interaction, you need to have a way to identify them. They are distinguished by the model already fit when the term is entered. That is, by the model consisting of the terms entered before the term in question.

In a three way design with factors A, B and C, there are 5 possible SS_C:

Sum of squares	Model after including C
SS(C|1)	$\mu + \gamma_k$ ("y=c")
SS(C|1,A)	$\mu + \alpha_i + \gamma_k$ ("y=a+c")
SS(C|1,B)	$\mu + \beta_j + \gamma_k$ ("y=b+c")
SS(C|1,A,B)	$\mu + \alpha_i + \beta_j + \gamma_k$ ("y=a+b+c")
SS(C|1,A,B,AB)	$\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k$ ("y=a+b+a.b+c")

There is some controversy as to what are the appropriate sums of squares in unbalanced cases.

MacAnova in general follows the principle that you usually should be fitting **hierarchical models**.

These are models that have the property that if a particular interaction is in the model, then all terms and main effects "contained" in it should also be in the model.

Example: In a 4 factor experiment, if you need the ABD interaction then you should keep A, B, D, AB, AD and BD in the model, even if they don't appear to be significant.

In fact, the way MacAnova does its computations, it enforces this. If you include an "including" interaction before an "included" term, the interaction SS has already "swept" up the included SS leaving nothing for the later term.

```
Cmd>  anova("y1=weeks + weeks.water + water")
Model used is y1=weeks + weeks.water + water
WARNING: cases with missing values deleted
WARNING: summaries are sequential
                DF          SS          MS
CONSTANT         1      5938.8      5938.8
weeks            4      1333.1      333.27
weeks.water      5      1372.6      274.53
water            0           0   undefined
ERROR1          19      1186.5      62.447
```

The SS for `weeks.water` is the sum of the actual interaction term and SS_{water}. The same is true of the DF: 5 = (5-1)(2-1) + 1. There are no model DF or SS for SS_{water} once the interaction is in the model.

```
Cmd> data <- read("","exmpl8.10")
exmpl8.10     96     4
) A data set from Oehlert (2000) \emph{A First Course in Design
) and Analysis of Experiments}, New York: W. H. Freeman.
)
) Data originally from Table 22 of Bruce Orman (1986) "Maize
) Germination and Seedling Growth at Suboptimal Temperatures",
) MS Thesis, University of Minnesota, St. Paul, MN.
)
) Table 8.9, p. 194
) Amylase activity in sprouted maize under various conditions.
) Column 1 is the temperature at which the assay takes place
). Levels 1 through 8 represent 40, 35, 30, 25, 20, 15, 13, and
) 10 degrees C.
) Column 2 is the growth temperature of the sprouts. Level 1 is
) 25  degrees, level 2 is 13 degrees.
) Column 3 is the variety of maize. Level 1 is B73, level 2 is
) Oh43.
) Column 4 is the amylase specific activity in international
units.
Read from file "TP1:Stat5303:Data:OeCh08.dat"

Cmd> makecols(data,assaytemp,growthtemp,variety,activity)

Cmd> assaytemp <- factor(assaytemp) # factor A

Cmd> growthtemp <- factor(growthtemp) # factor B

Cmd> variety <- factor(variety) # factor C

Cmd> list(assaytemp,growthtemp,variety,activity)
activity        REAL   96
assaytemp       REAL   96    FACTOR with 8 levels
growthtemp      REAL   96    FACTOR with 2 levels
variety         REAL   96    FACTOR with 2 levels
```

Make the data unbalanced by replacing the first case with MISSING.

```
Cmd> activity[1] <- ? # or activity[1] <- NA

Cmd> hconcat(assaytemp,growthtemp,variety)[1,]
(1,1)          1           1           1
```


This is best analyzed in terms of logs:

```
Cmd> logy <- log(activity)

Cmd> anova("logy=(assaytemp + growthtemp + variety)^3",fstat:T)
Model used is logy=(assaytemp + growthtemp + variety)^3
WARNING: cases with missing values deleted
WARNING: summaries are sequential
             DF          SS          MS           F     P-value
CONSTANT      1      3200.5      3200.5   6.012e+05           0
assaytemp     7      3.0628     0.43755    82.19202           0
growthtemp    1    0.001396    0.001396     0.26223     0.61038
variety       1     0.55282     0.55282   103.84598  5.9679e-15
assaytemp.
 growthtemp   7     0.06407   0.0091529     1.71935     0.12055
assaytemp.
 variety      7    0.025892   0.0036989     0.69483     0.67608
growthtemp.
 variety      1    0.078632    0.078632    14.77084  0.00028496
assaytemp.
 growthtemp.
 variety      7    0.053554   0.0076506     1.43715     0.20654
ERROR1       63     0.33538   0.0053235
```

There is no problem testing the ABC interaction since it is the last term. It is not significant.

You can also test BC since it is the last two-factor interaction. Its SS is $SS(BC \mid 1,A,B,C,AB,AC)$ and is significant.

But you can't test AB or AC from these sums of squares since their SS do not follow BC.

Find SS(AC | 1,A,B,AB,BC)

```
Cmd> anova("logy=assaytemp + growthtemp + variety +\
growthtemp.variety + assaytemp.growthtemp +
assaytemp.variety",\
fstat:T)
Model used is logy=assaytemp + growthtemp + variety +\
growthtemp.variety + assaytemp.growthtemp + assaytemp.variety
WARNING: cases with missing values deleted
WARNING: summaries are sequential
              DF         SS         MS            F     P-value
CONSTANT       1     3200.5     3200.5  5.7602e+05 8.6928e-139
assaytemp      7     3.0628    0.43755    78.74947  1.2012e-30
growthtemp     1   0.001396   0.001396     0.25125     0.61777
variety        1    0.55282    0.55282    99.49646  4.4379e-15
growthtemp.
 variety       1   0.075538   0.075538    13.59537  0.00044398
assaytemp.
 growthtemp    7   0.067028  0.0095754     1.72337     0.11756
assaytemp.
 variety       7   0.026029  0.0037184     0.66924     0.69725
ERROR1        70    0.38893  0.0055562
```

Find SS(AB | 1,A,B,AC,BC)

```
Cmd> anova("logy=assaytemp + growthtemp + variety +\
growthtemp.variety+assaytemp.variety+assaytemp.growthtemp",\
fstat:T)
Model used is logy=assaytemp + growthtemp + variety +\
growthtemp.variety + assaytemp.variety + assaytemp.growthtemp
WARNING: cases with missing values deleted
WARNING: summaries are sequential
              DF         SS         MS            F     P-value
CONSTANT       1     3200.5     3200.5  5.7602e+05 8.6928e-139
assaytemp      7     3.0628    0.43755    78.74947  1.2012e-30
growthtemp     1   0.001396   0.001396     0.25125     0.61777
variety        1    0.55282    0.55282    99.49646  4.4379e-15
growthtemp.
 variety       1   0.075538   0.075538    13.59537  0.00044398
assaytemp.
 variety       7     0.0259  0.0037001     0.66593     0.69998
assaytemp.
 growthtemp    7   0.067156  0.0095937     1.72668     0.11679
ERROR1        70    0.38893  0.0055562
```