Stat 512 Class 32

- Pooling SS
- Strategy for analysis or two-way studies
  - Interaction is not significant
  - Interaction is significant
- Quantitative factors
- One observation per cell

Pooling SS

- Data = Model + Residual
- When we remove a term from the `model', we put this variation and the associated df into `residual'
- This is called pooling
- A benefit is that we have more df for error and a simpler model

Pooling SSE and SSAB

- For model with interaction
  - SSAB=24, dfAB=2
  - SSE=62, dfE=6
  - MSE=10.33
- For the model with main effects only
  - SSE=62+24=86, dfE=6+2=8
  - MSE=10.75

An analytical strategy

- Run the model with main effects and the two-way interaction
- Plot the data, the means and look at the normal quantile plot
- Check the significance test for the interaction

AB interaction ns

- Rerun the analysis without the interaction
- For a main effect that is not significant
  - There is no evidence to conclude that the levels of this explanatory variable are associated with different means of the response variable
  - Model could be rerun without this factor giving a one-way anova

AB interaction ns (2)

- If both main effects are not significant
  we may use a one population model
- For a main effect with more than two levels that is significant, use the tukey multiple comparison procedure

AB interaction ns (3)

- Marginal means for the significant main effects and overall means can be estimated
  - Use the table.means and TukeyHSD functions

AB interaction sig but not important

- Plots and a careful examination of the cell means may indicate that the interaction is not very important even though it is statistically significant
- Use the marginal means for each significant main effect to describe the important results
- You may need to qualify these results using the interaction

AB interaction sig but not important (2)

- Use the methods described for the situation where the interaction is not significant but keep the interaction in the model
- Carefully interpret the marginal means as averages over the levels of the other factor

AB interaction is sig and important

- Options include
  - Treat as a one-way with IJ levels; use tukey to compare means;
  - Report that the interaction is significant; plot the means and describe the pattern
  - Analyze the levels of A for each level of B or vice versa

One Quantitative factor and one categorical

- Plot the means vs the quantitative factor for each level of the categorical factor
- Consider linear and quadratic terms for the quantitative factor
- Consider different slopes for the different levels of the categorical factor
- Lack of fit analysis can be useful

Two Quantitative factors

- Plot the means
  - vs A for each level B
  - vs B for each level A
- Consider linear and quadratic terms
- Consider products to allow for interaction
- Lack of fit analysis can be useful

One observation per cell

- For Y_{ijk} we use
  - i to denote the level of the factor A
  - j to denote the level of the factor B
  - k to denote the k^{th} observation in cell (i,j)
- i = 1, . . . , I levels of factor A
- j = 1, . . . , J levels of factor B
- k = 1 observation in cell (i,j) (K=1)

Factor effects model

- $\mu_{ij} = \mu + \alpha_i + \beta_j$
- μ is the overall mean
- α_i is the main effect of A
- β_j is the main effect of B
- Because we have only one observation per cell, we do not have enough information to estimate the interaction in the usual way

Constraints

- $\Sigma_i \alpha_i = 0$
- $\Sigma_j \beta_j = 0$

Estimates for Factor effects model

- $\hat{\mu} = Y_{...} = (\Sigma_{ij} Y_{ij1})/(IJ)$
- $\hat{\mu}_{i.} = Y_{i..} = (\Sigma_j Y_{ij1})/J$
- $\hat{\mu}_{.j} = Y_{.j.} = (\Sigma_i Y_{ij1})/I$
- $\hat{\alpha}_i = \hat{\mu}_{i.} - \hat{\mu} = Y_{i..} - Y_{...}$
- $\hat{\beta}_j = \hat{\mu}_{.j} - \hat{\mu} = Y_{.j.} - Y_{...}$

R LM Constraints

- $\alpha_1 = 0$ (1 constraint)
- $\beta_1 = 0$ (1 constraint)

ANOVA Table

Source	df	SS	MS	F
A	I-1	SSA	MSA	MSA/MSE
B	J-1	SSB	MSB	MSB/MSE
Error	(I-1)(J-1)	SSE	MSE	
Total	IJ-1	SST	MST	

Expected Mean Squares

- E(MSE) = σ^2
- E(MSA) = $\sigma^2 + J(\Sigma_i \alpha_i^2)/(I-1)$
- E(MSB) = $\sigma^2 + I(\Sigma_j \beta_j^2)/(J-1)$
- Here, α_i and β_j, are defined with the usual factor effects constraints

KNNL Example

- KNN p 882
- Y is the premium for auto insurance
- A is the size of the city, I=3 levels: small, medium and large
- B is the region, J=2: East, West
- K=1 the response is the premium charged by a particular company

The data

```
ins<-read.table('ch20ta02.txt',
col.names=c("premium", "size",
"region"));
ins$sizea<-rep("1_small",6);
ins$sizea[ins$size==2]<-"2_medium";
ins$sizea[ins$size==3]<-"3_large";
```

Ins

```
  premium size region    sizea
1     140    1      1  1_small
2     100    1      2  1_small
3     210    2      1 2_medium
4     180    2      2 2_medium
5     220    3      1  3_large
6     200    3      2  3_large
```

aov and lm

```
ins$sizea<-factor(ins$sizea);
ins$region<-factor(ins$region);
obj<-aov(premium~sizea+region,
ins);
summary(obj)
model.tables(obj, 'means')
TukeyHSD(obj)
obj2<-lm(premium~size+region,ins);
summary(obj2)
predict.lm(obj2)
```

Output (lm)

```
F-statistic: 71 on 3 and 2 DF,
 p-value: 0.01392
```

Output (aov)

```
Df Sum Sq Mean Sq F value  Pr(>F)
sizea   2 9300 4650 93 0.01064 *
region  1 1350 1350 27 0.03510 *
Res     2  100   50
```

Output lm

```
       Est Std t value Pr(>|t|)
Int  135.0 5.8  23.383  0.00182
sz2   75.0 7.1  10.607  0.00877
sz3   90.0 7.1  12.728  0.00612
rg2  -30.0 5.8  -5.196  0.03510
```

Check vs predicted

```
reg sizea muhat
1   1_s    135
2   1_s    105=135-30
1   2_m    210=135+75
2   2_m    180=135+75-30
1   3_l    225=135+90
2   3_l    195=135+90-30
```

Multiple comparisons Size

```
$sizea   diff lwr  upr     p adj
2_m-1_s   75  33.3 116.65 0.016
3_l-1_s   90  48.3 131.65 0.011
3_l-2_m   15 -26.7  56.65 0.288
```

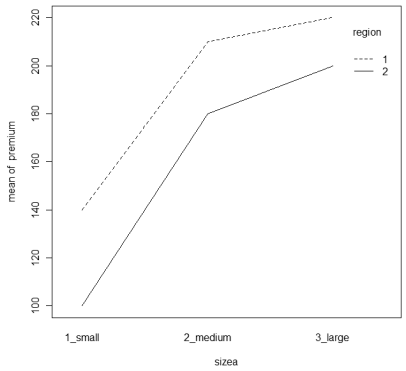
Multiple comparisons Region

```
    diff  lwr   upr     p adj
2-1 -30 -54.82 -5.18 0.0350019
```

The anova results told us that these were different

Plot the means

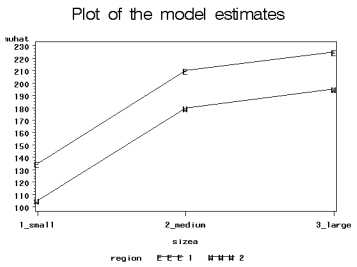
```
with(ins,interaction.plot
(sizea, region, premium))
```



Plot the estimated model

```
symbol1 v='E' i=join c=blue;
symbol2 v='W' i=join c=green;
title1 'Plot of the model
            estimates';
proc gplot data=a2;
   plot muhat*sizea=
          region;
run;
```

The plot



Tukey test for additivity

- One additional term is added to the model (θ)
- $\mu_{ij} = \mu + \alpha_i + \beta_j + \theta\alpha_i\beta_j$
- We use one degree of freedom to estimate θ
- There are other variations, eg $\theta_i\beta_j$
- tukey.add.test {asbio}