Lecture 26

Basics of Two-Way ANOVA

STAT 512

Spring 2011

Background Reading

KNNL: Chapter 19

Topic Overview

- Two-way ANOVA Models
- Main Effects; Interaction
- Analysis of Variance Table / Tests

Two-way ANOVA

- Response variable Y_{ijk} is continuous
- Have two categorical explanatory variables (call them **Factor A** and **Factor B**)
- Factor A has levels $i = 1$ to a
- Factor B has levels $j = 1$ to b
- Each combination of levels (i,j) labels the ***treatment combination*** or cell.

Two-way ANOVA (2)

- A third subscript k indicates observation number in cell (i,j). $k = 1$ to n_{ij} (for now assume **balanced design**; equal sample sizes with $n_{ij} \equiv n$)
- Could analyze as a one-way ANOVA by taking each (i,j) combination as a different level of a single factor.

Cash Offers Example

- In addition to AGE, consider GENDER as a second factor.
- a = 3 levels of age (young, middle, elderly)
- b = 2 levels of gender (female, male)
- n = 6 observations per age*gender combination (total 36 observations)

Cell Means Model

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

where $\varepsilon_{ijk} \sim N\left(0, \sigma^2\right)$ are independent

- Estimate μ_{ij} by cell mean $\bar{Y}_{ij\bullet}$
- Estimate factor level means (mean for one level of given factor across all levels of other factor), as follows:

$$\hat{\mu}_{i\bullet} = \bar{Y}_{i\bullet\bullet} \quad \text{and} \quad \hat{\mu}_{\bullet j} = \bar{Y}_{\bullet j\bullet}$$

- Estimate grand mean by $\hat{\mu} = \bar{Y}_{\bullet\bullet\bullet}$
- Disadvantage – need contrasts to separate effects.

Factor Effects Model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where $\varepsilon_{ijk} \sim N(0, \sigma^2)$ are independent

and $\sum \alpha_i = \sum \beta_i = \sum (\alpha\beta)_{ij} = 0$

- Constraints required to keep model from being over-parameterized
- Advantage: Effects can be analyzed separately. This is the model we want to use.

Factor Effects Model (2)

- **Grand Mean**: Estimate μ by $\bar{Y}_{\bullet\bullet\bullet}$
- **Main Effects**
  - Estimate α_i by $\hat{\alpha}_i = \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet}$
  - Estimate β_i by $\hat{\beta}_j = \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}$
- **Interaction**:
  - Estimate $(\alpha\beta)_{ij}$ by

$$\widehat{(\alpha\beta)}_{ij} = \bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet}$$

  - If these are zero, effects are ***additive***.

Cash Offers Example

- SAS Code: cashoffers_twoway.sas
- MEANS procedure can be used to get the estimates.

```
proc sort data=cash;
  by gender age;
proc means;
  class gender age;
  var offer;
  output out=means mean = mean_offer;
proc print;
run;
```

Output

Obs	gender	age	_TYPE_	_FREQ_	mean
1			0	36	23.5556
2		Elderly	1	12	21.4167
3		Middle	1	12	27.7500
4		Young	1	12	21.5000
5	Female		2	18	23.1667
6	Male		2	18	23.9444
7	Female	Elderly	3	6	20.5000
8	Female	Middle	3	6	27.6667
9	Female	Young	3	6	21.3333
10	Male	Elderly	3	6	22.3333
11	Male	Middle	3	6	27.8333
12	Male	Young	3	6	21.6667

Estimates

- Lines 7-12 contain the estimates for the cell means model.
- Can construct estimates for the factor effects model from this table
- Example:

$$\hat{\alpha}_{female} = 23.1667 - 23.5556 = -0.3889$$

$$\left(\widehat{\alpha\beta}\right)_{male,young} = 21.67 - 23.94 - 21.5 + 23.56 = -0.21$$

Summary of Estimates

$$\hat{\alpha}_{male} = 0.3889$$

$$\hat{\alpha}_{female} = -0.3889$$

$$\hat{\beta}_{eld} = -2.14$$

$$\hat{\beta}_{mid} = 4.58$$

$$\hat{\beta}_{yng} = -2.44$$

$$\left(\widehat{\alpha\beta}\right)_{m,e} = 0.53$$

$$\left(\widehat{\alpha\beta}\right)_{m,m} = -0.31$$

$$\left(\widehat{\alpha\beta}\right)_{m,y} = -0.22$$

$$\left(\widehat{\alpha\beta}\right)_{f,e} = -0.53$$

$$\left(\widehat{\alpha\beta}\right)_{f,m} = 0.31$$

$$\left(\widehat{\alpha\beta}\right)_{f,y} = 0.22$$

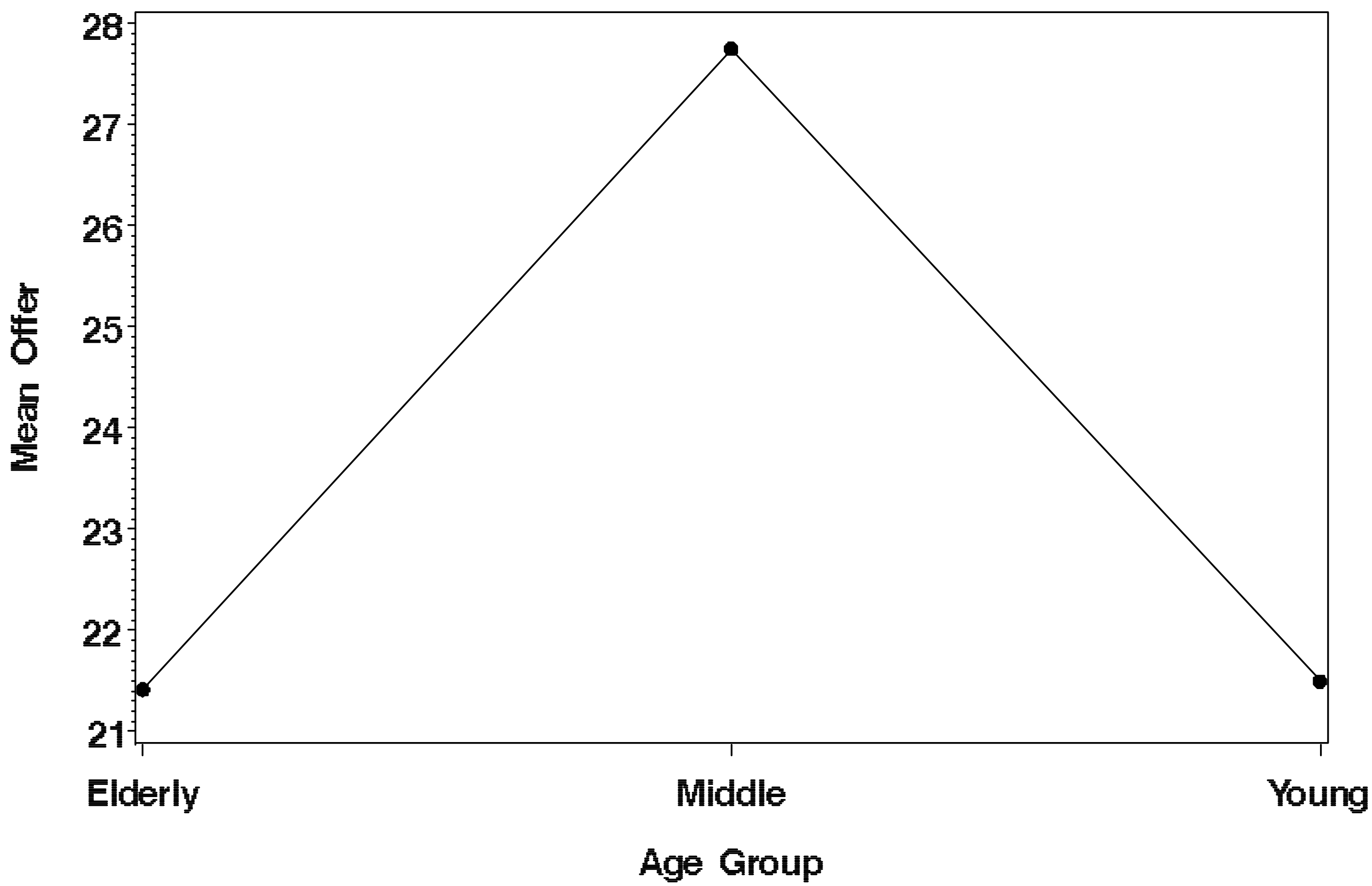
Summary of Estimates (2)

- Largest effect is Main Effect for AGE
- Effect of Gender is small compared to effect for Age, but should look at interaction
- We haven't yet looked at "significance" – sizes of effects relative to standard errors to determine if they are significant.
- Can look at these effects in a plot too! Visual representation is often more appealing and informative.

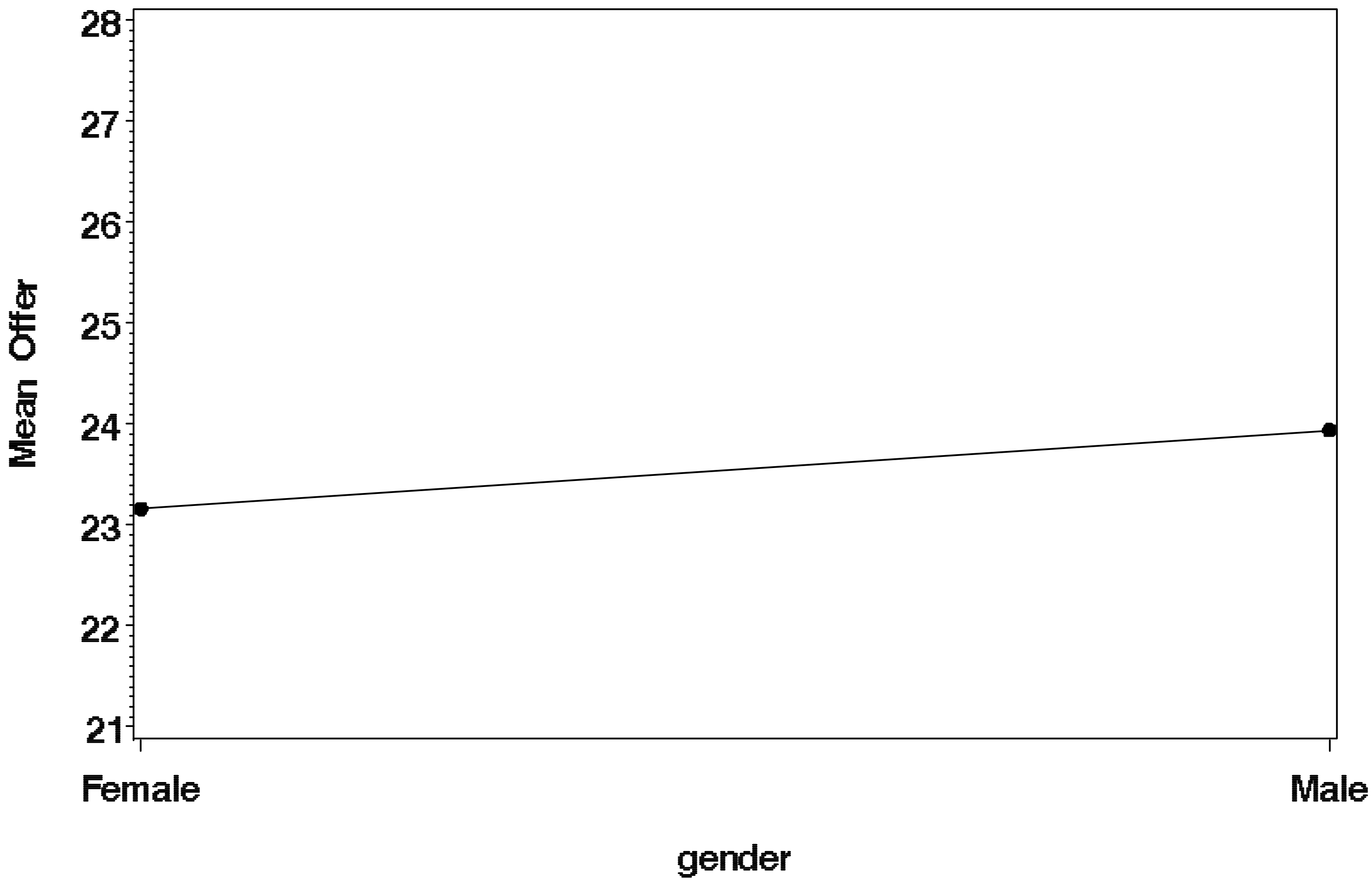
Main Effects Plot

```
data a1; set means; if _TYPE_=1;
data a2; set means; if _Type_=2;

symbol1 v=dot i=join;
axis1 label=(angle=90 'Mean Offer')
    order = 21 to 28 by 1;

proc gplot data=a1;
  plot mean_offer*age /vaxis=axis1;
proc gplot data=a2;
  plot mean_offer*gender /vaxis=axis1;
```

Main Effects Plot (Age)



Main Effects Plot (Gender)



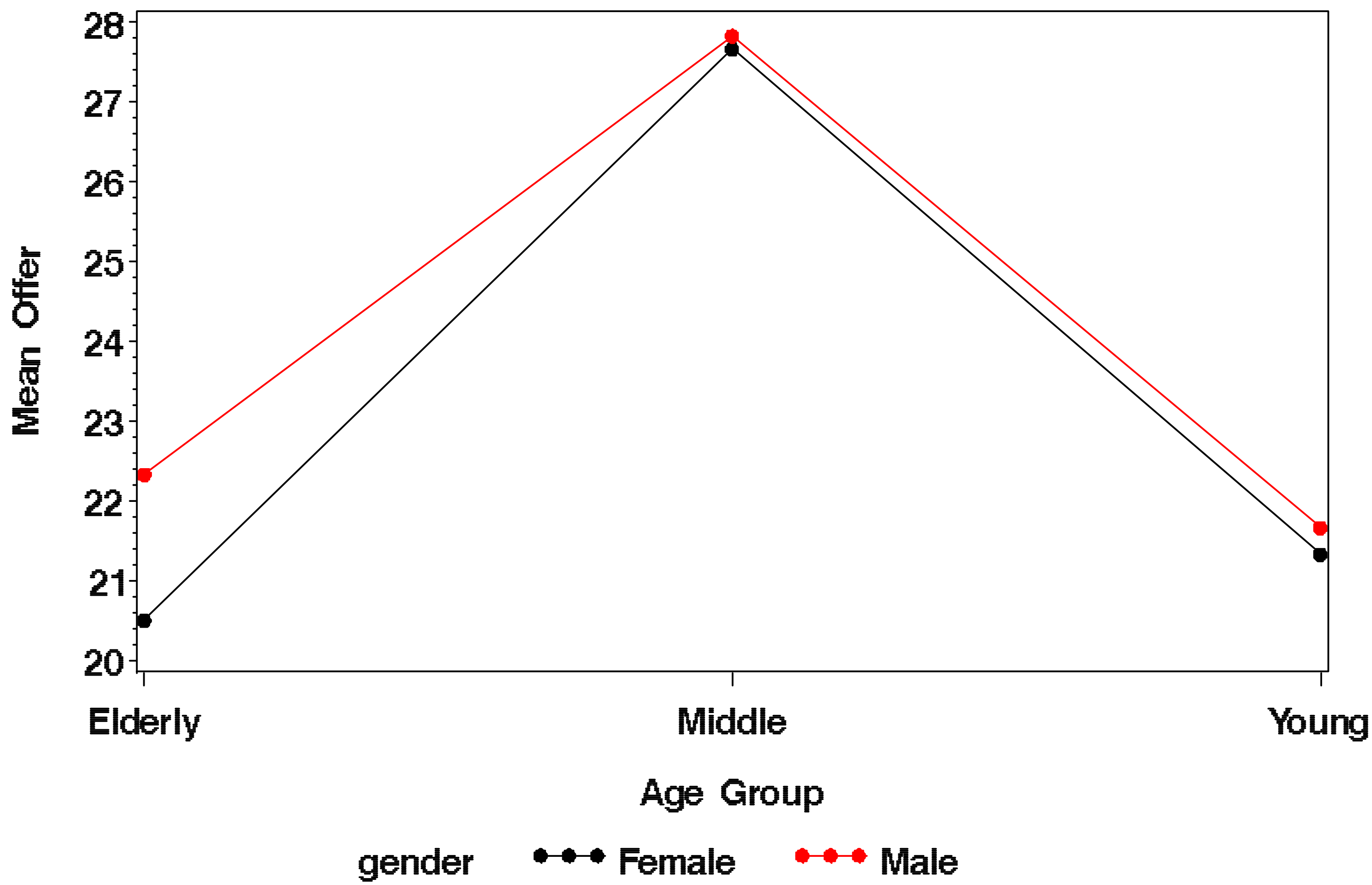
Interaction Plot

- Plot means against levels of one factor, with different lines for the other factor

```
data a3; set means; if _Type_=3;
axis1 label=(angle=90 'Mean Offer');

proc gplot data=a3;
  plot mean_offer*age=gender
                /vaxis=axis1;
```

- Parallel Lines indicate **additive model** (no interaction present)

Interaction Plot



Analysis of Variance Table

- Model line treats as one-way ANOVA and does not separate the effects.
- Type I / Type III SS can be used to investigate interaction and main effects
- Often replace model line by Type I SS to form "expanded" ANOVA Table
- For **balanced design**, Type I / Type III SS are the same.

Analysis of Variance Table (2)

- Model SS gets partitioned into SSA, SSB, and SSAB.
- Associated degrees of freedom are $a - 1$, $b - 1$, and $(a - 1)(b - 1)$
- Error degrees of freedom calculated by subtracting everything else from total.

Example

- 4 levels of factor A, 3 levels of factor B
- 6 observations per cell

SOURCE	DF
A	3
B	2
A*B	6
Error	60
Total	71

F-tests

- Interaction:

$$H_0 : all\,(\alpha\beta)_{ij} = 0 \text{ vs. } H_a : not\ all\,(\alpha\beta)_{ij}\ equal\,0$$

- Main Effect of Factor A:

$$H_0 :\ \alpha_1 = \alpha_2 = ... = \alpha_a = 0 \text{ vs. } H_a : not\ all\,\alpha_i\ equal\,0$$

- Main Effect of Factor B:

$$H_0 :\ \beta_1 = \beta_2 = ... = \beta_b = 0 \text{ vs. } H_a : not\ all\,\beta_i\ equal\,0$$

F-tests

- Based on Expected Mean Squares (see pages 840-841); Mean Squares calculated as usual (SS / DF)
- When the effects are ***fixed***,
  - Ratio of MSAB / MSE tests for interaction effect (***test this first since interpretation of main effects depend on significance of interaction)***.
  - Ratio of MSA / MSE tests for factor A main effect
  - Ratio of MSB / MSE tests for factor B main effect.

Cash Offers

```
proc glm data=cash;
  class age gender;
  model offer=age gender age*gender;

proc glm data=cash;
  class age gender;
  model offer=age|gender;
```

- Two ways to write the same model in SAS.
- Having an interaction means both factors are important. So we would never use a model that just has interaction without including main effects.

ANOVA Results

Source	DF	SS	MS	F Value	Pr > F
age	2	316.72	158.36	66.29	<.0001
gender	1	5.44	5.44	2.28	0.1416
age*gender	2	5.06	2.53	1.06	0.3597
Error	30	71.67	2.39		
Total	35	398.89			

- Interaction Effect is <u>not</u> significant; proceed to test main effects.
- Gender Effect is <u>not</u> significant
- Age Effect is significant

Castle Bakery Co. Example

- Experimental study designed to examine the effect of shelf height (bottom, middle, top) and shelf width (regular, wide) on the sales of bread (measured in cases sold).
- Twelve stores studied, six treatments randomly assigned to two stores each
- Data in Table 19.7; SAS code in bakery.sas
- Define A = height, B = width

Steps in Analysis

1. Check some basic plots. Examine ANOVA and check assumptions.
2. Does interaction appear to be important?
   - If yes, must analyze on the interaction level and may not be able to look at main effects.
   - If no, analyze main effects.
3. Draw appropriate conclusions from ANOVA and plots.
4. Summarize results.

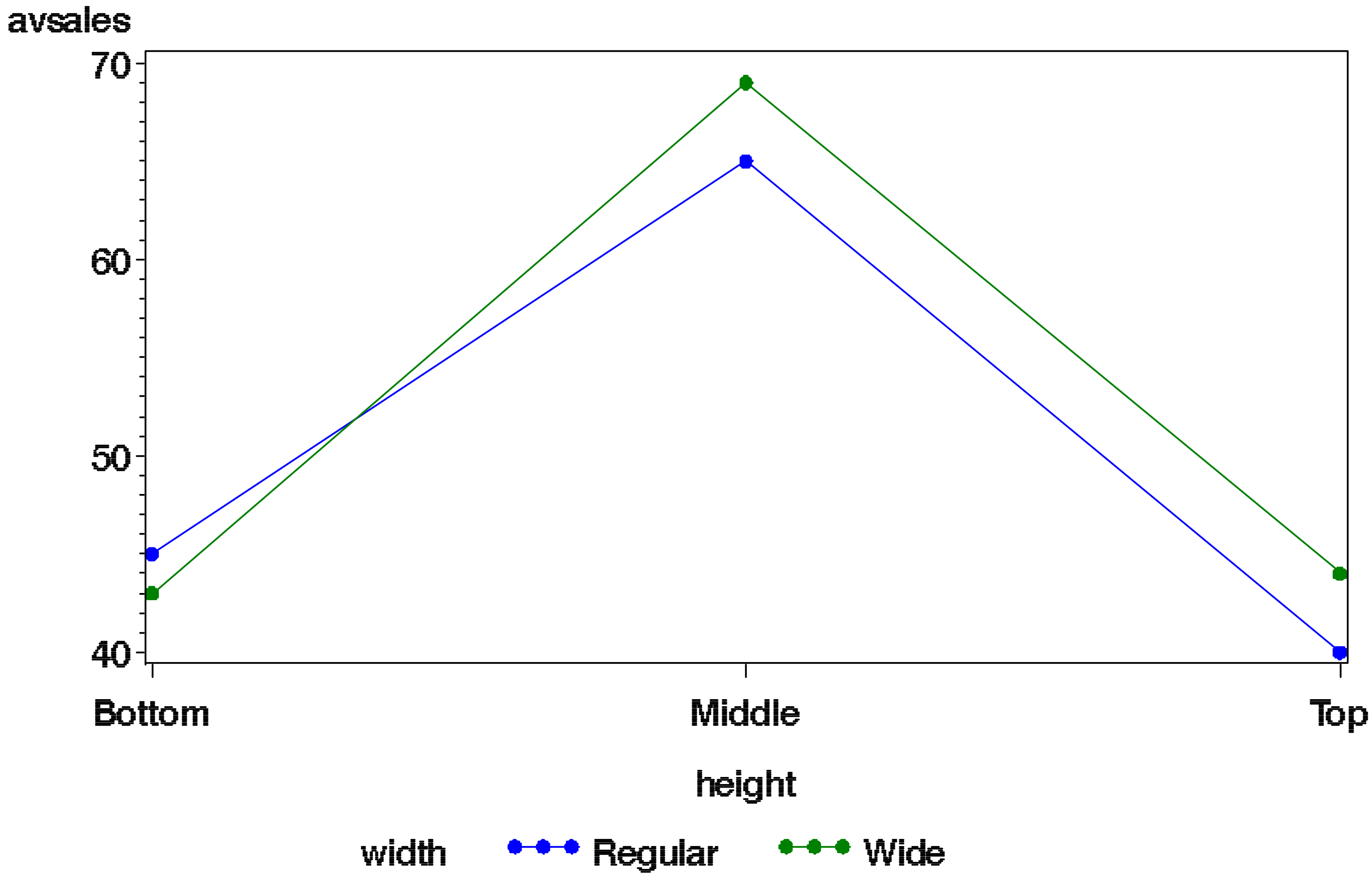
Interaction Plot

- Allows for assessment of the sizes for interaction effects and main effects
- Get cell means from PROC MEANS
- Two possible plots
  - Plot means against Height, with different lines representing different Widths
  - Plot means against Width, with different lines representing different heights

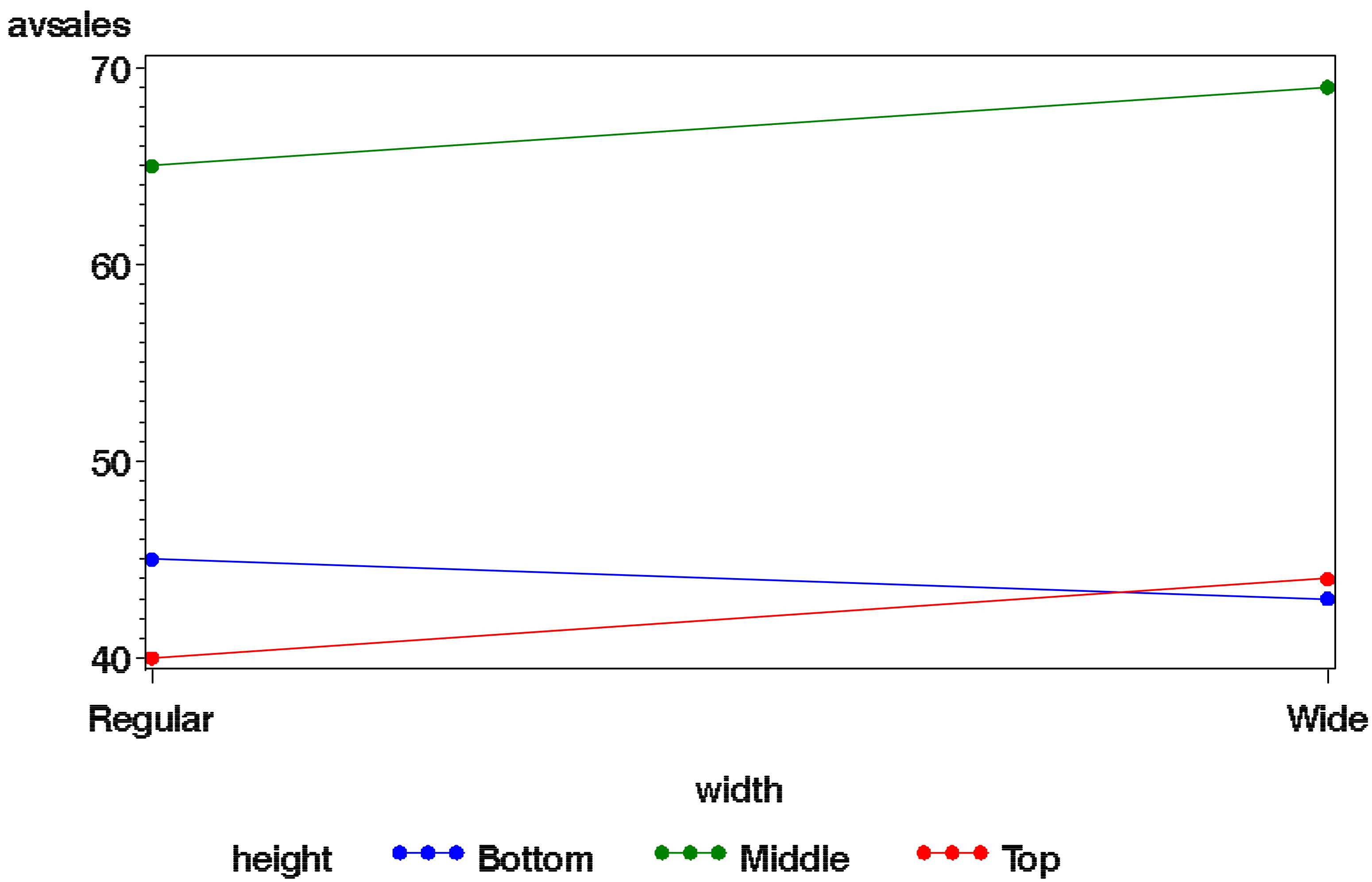
Interaction Plot

```
proc means data=bakery;
   var sales;
   by height width;
   output out=means mean=avsales;
proc print data=means;

symbol1 v=dot i=join c=blue;
symbol2 v=dot i=join c=green;
symbol3 v=dot i=join c=red;
proc gplot data=means;
   plot avsales*height=width;
proc gplot data=means;
   plot avsales*width=height;
```

Interaction Plot



Interaction Plot



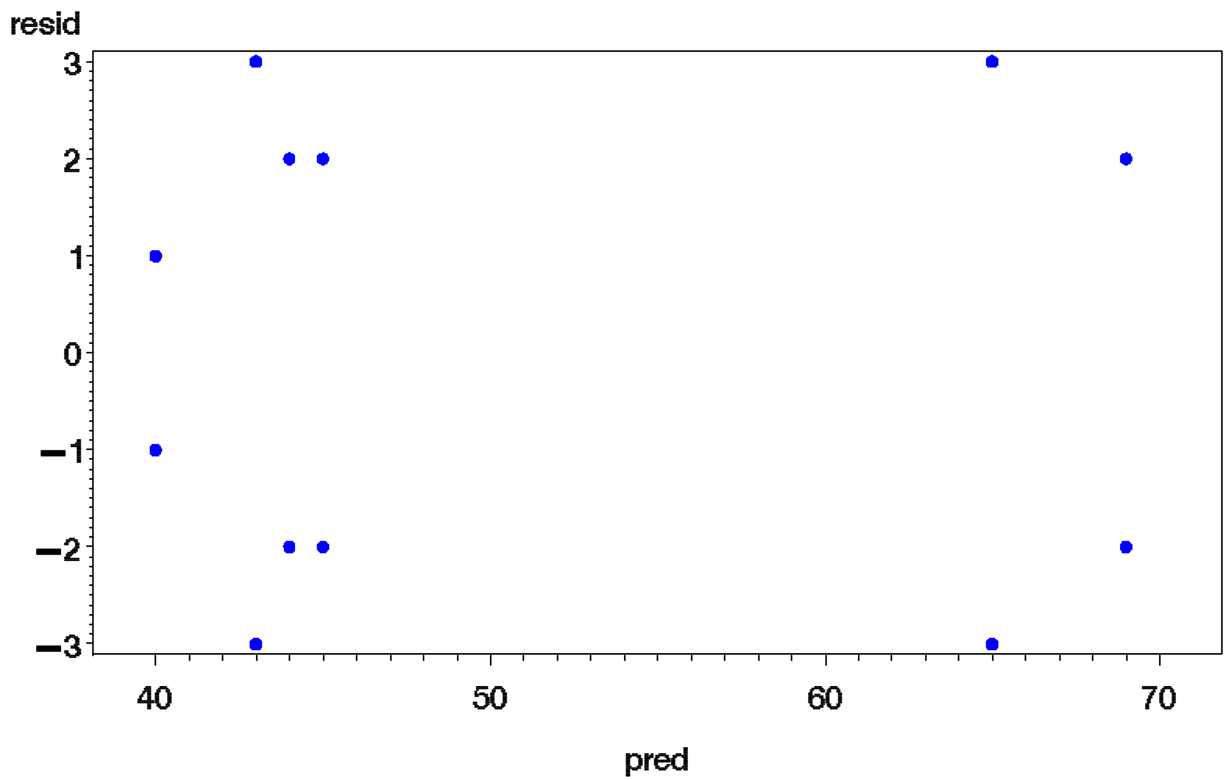
Questions

- Interaction?

- Does height of display affect sales?

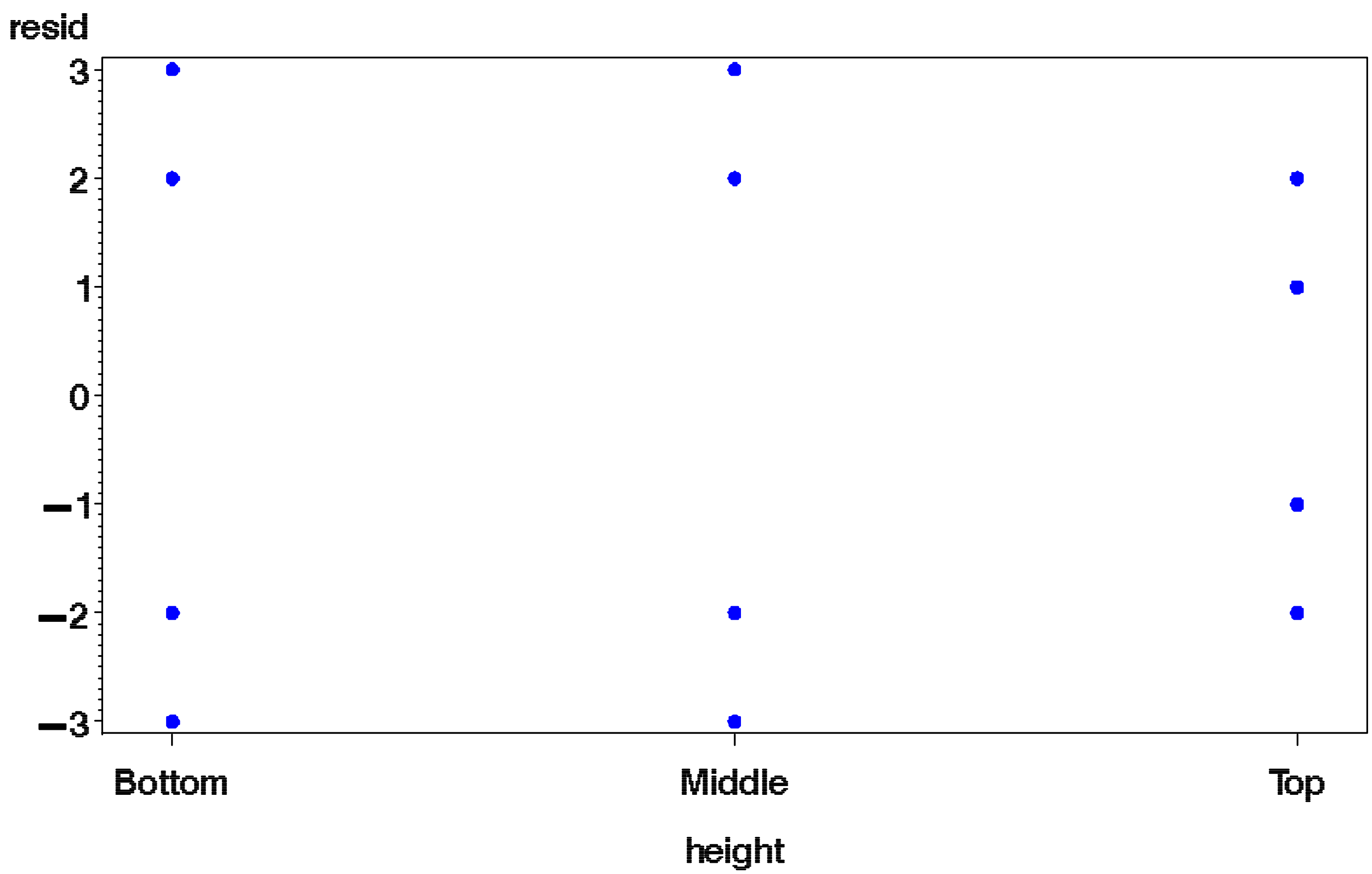
- Does width of display affect sales?

Check Assumptions



- Residuals vs Predicted Values – no evident problems

Check Assumptions

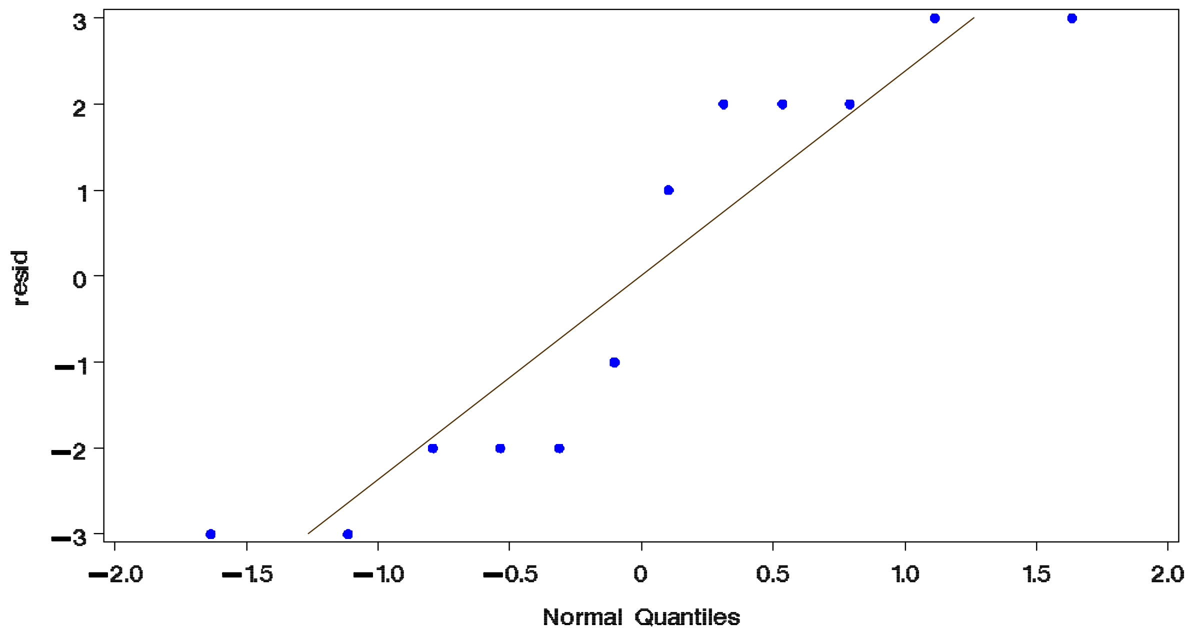


- Residuals vs Height – no evident problems

Check Assumptions

- Residuals vs Width – no evident problems

Check Assumptions



- Perhaps not quite normal, but remember that ANOVA is pretty robust to departures from normality.

ANOVA

```
proc glm data=bakery;
   class height width;
   model sales=height width height*width;
   lsmeans height width height*width
           /adjust=tukey cl tdiff pdiff;
```

Source	DF	SS	MS	F Value	Pr > F
height	2	1544	772	74.71	<.0001
width	1	12	12	1.16	0.3226
height*width	2	24	12	1.16	0.3747
Error	6	62	10.3		
Total	11	1642			

LSMEANS Output

- LSMEANS statement used when multiple factors
- Means are compared after "adjusting" for the levels of the other factor

```
The GLM Procedure Least Squares Means
Adjustment for Multiple Comparisons: Tukey
                                  LSMEAN
height    sales LSMEAN            Number
Bottom      44.0000000                 1
Middle      67.0000000                 2
Top         42.0000000                 3
```

LSMEANS Output

```
 Least Squares Means for Effect height
t for H0: LSMean(i)=LSMean(j) / Pr > |t|

      Dependent Variable: sales
```

i/j	1	2	3
1		-10.1187	0.879883
		0.0001	0.6714
2	10.11865		10.99853
	0.0001		<.0001
3	-0.87988	-10.9985	
	0.6714	<.0001	

LSMEANS Output

height	sales LSMEAN	95% Conf. Limits	
Bottom	44.00	40.067	47.933
Middle	67.00	63.067	70.933
Top	42.00	38.067	45.933

- For completely balanced design, the results will be the same as a MEANS statement, but get into the habit of using LSMEANS because sometimes we aren't so lucky as to have a completely balanced design

LSMEANS Output

- Can combine LSMEANS output by hand to produce chart similar to “LINES” chart

height	sales LSMEAN	Tukey Group
Middle	67.00	A
Bottom	44.00	B
Top	42.00	B

- Middle shelf has significantly higher sales

Upcoming in Lecture 27...

- More on Two-way ANOVA (Chapter 19)
- Focus on Interactions