Statistics 512: Applied Linear Models

Topic 7

Topic Overview

This topic will cover

- Two-Way Analysis of Variance (ANOVA) (§19)
- Interactions (§19)
- One Case Per Treatment (§20)

Chapter 19: Two-way ANOVA

The response variable Y is continuous.

There are now *two* categorical explanatory variables (factors). Call them factor A and factor B instead of X_1 and X_2. (We will have enough subscripts as it is!)

Data for Two-way ANOVA

- Y, the response variable
- Factor A with levels $i = 1$ to a
- Factor B with levels $j = 1$ to b
- A particular *combination* of levels is called a treatment or a cell. There are ab treatments.
- $Y_{i,j,k}$ is the kth observation for treatment (i, j), $k = 1$ to n

In Chapter 19, we for now assume equal sample size in each treatment combination ($n_{i,j} = n > 1$; $n_T = abn$). This is called a *balanced design*. In later chapters we will deal with unequal sample sizes, but it is more complicated.

Notation

For $Y_{i,j,k}$ the subscripts are interpreted as follows:

- i denotes the level of the factor A
- j denotes the level of the factor B

- k denotes the kth observation in cell or treatment (i, j)

$i = 1, \ldots, a$ levels of factor A
$j = 1, \ldots, b$ levels of factor B
$k = 1, \ldots, n$ observations in cell (i, j)

KNNL Example

- KNNL page 832 (`nknw817.sas`)
- response Y is the number of cases of bread sold.
- factor A is the height of the shelf display; $a = 3$ levels: bottom, middle, top.
- factor B is the width of the shelf display; $b = 2$ levels: regular, wide.
- $n = 2$ stores for each of the $3 \times 2 = 6$ treatment combinations ($n_T = 12$)

Read the data

```
data bread;
        infile 'h:\System\Desktop\CH19TA07.DAT';
        input sales height width;
proc print data=bread;
```

Obs	sales	height	width
1	47	1	1
2	43	1	1
3	46	1	2
4	40	1	2
5	62	2	1
6	68	2	1
7	67	2	2
8	71	2	2
9	41	3	1
10	39	3	1
11	42	3	2
12	46	3	2

Model Assumptions

We assume that the response variable observations are independent, and normally distributed with a mean *that may depend on the levels of the factors A and B*, and a variance that does not (is constant).

Cell Means Model

$Y_{i,j,k} = \mu_{i,j} + \epsilon_{i,j,k}$ where

- $\mu_{i,j}$ is the theoretical mean or expected value of all observations in cell (i, j).
- the $\epsilon_{i,j,k}$ are iid $N(0, \sigma^2)$
- $Y_{i,j,k} \sim N(\mu_{i,j}, \sigma^2)$, independent

There are $ab + 1$ parameters of the model: $\mu_{i,j}$, for $i = 1$ to a and $j = 1$ to b; and σ^2.

Parameter Estimates

- Estimate $\mu_{i,j}$ by the mean of the observations in cell (i, j), $\bar{Y}_{i,j.} = \frac{\sum_k Y_{i,j,k}}{n}$.
- For each (i, j) combination, we can get an estimate of the variance $\sigma^2_{i,j}$: $s^2_{i,j} = \frac{\sum_k (Y_{i,j,k} - \bar{Y}_{i,j.})^2}{n-1}$.
- Combine these to get an estimate of σ^2, since we assume they are all equal.
- In general we pool the $s^2_{i,j}$, using weights proportional to the df, $n_{i,j} - 1$.
- The pooled estimate is $s^2 = \frac{\sum_{i,j}(n_{i,j}-1)s^2_{i,j}}{\sum_{i,j}(n_{i,j}-1)} = \frac{\sum_{i,j}(n_{i,j}-1)s^2_{i,j}}{n_T - ab}$.
- Here, $n_{i,j} = n$, so $s^2 = \frac{\sum s^2_{i,j}}{ab} = MSE$.

Investigate with SAS

Note we are including an interaction term which is denoted as the product of A and B. It is not literally the product of the levels, but it would be if we used indicator variables and did regression. Using `proc reg` we would have had to create such a variable with a `data` step. In `proc glm` we can simply include `A*B` in the `model` statement, and it understands we want the interaction included.

```
proc glm data=bread;
   class height width;
   model sales=height width height*width;
   means height width height*width;

The GLM Procedure
   Class Level Information
Class         Levels    Values
height             3    1 2 3
width              2    1 2

Number of observations    12
```

means statement height

```
The GLM Procedure

Level of           ------------sales------------
height       N             Mean          Std Dev
1            4       44.0000000       3.16227766
2            4       67.0000000       3.74165739
3            4       42.0000000       2.94392029
```

means statement width

```
Level of           ------------sales------------
width        N             Mean          Std Dev
1            6       50.0000000       12.0664825
2            6       52.0000000       13.4313067
```

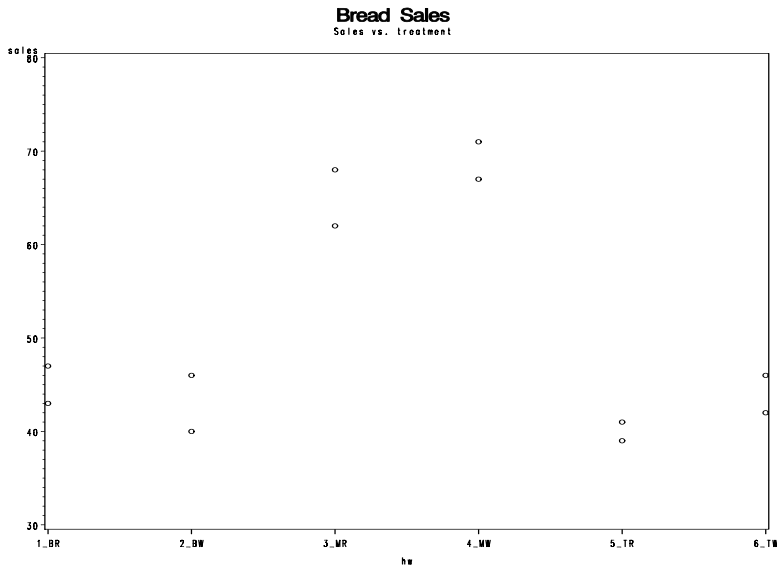
means statement height × width

```
Level of     Level of           ------------sales------------
height       width        N             Mean          Std Dev
1            1            2       45.0000000       2.82842712
1            2            2       43.0000000       4.24264069
2            1            2       65.0000000       4.24264069
2            2            2       69.0000000       2.82842712
3            1            2       40.0000000       1.41421356
3            2            2       44.0000000       2.82842712
```

Code the factor levels and plot

(We're just doing this for a nice plot; it is not necessary for the analysis.)

```
data bread;
set bread;
    if height eq 1 and width eq 1 then hw='1_BR';
    if height eq 1 and width eq 2 then hw='2_BW';
    if height eq 2 and width eq 1 then hw='3_MR';
    if height eq 2 and width eq 2 then hw='4_MW';
    if height eq 3 and width eq 1 then hw='5_TR';
    if height eq 3 and width eq 2 then hw='6_TW';
title2 'Sales vs. treatment';
symbol1 v=circle i=none;
proc gplot data=bread;
   plot sales*hw;
```



Put the means in a new dataset

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

Obs	height	width	_TYPE_	_FREQ_	avsales
1	1	1	0	2	45
2	1	2	0	2	43
3	2	1	0	2	65
4	2	2	0	2	69
5	3	1	0	2	40
6	3	2	0	2	44

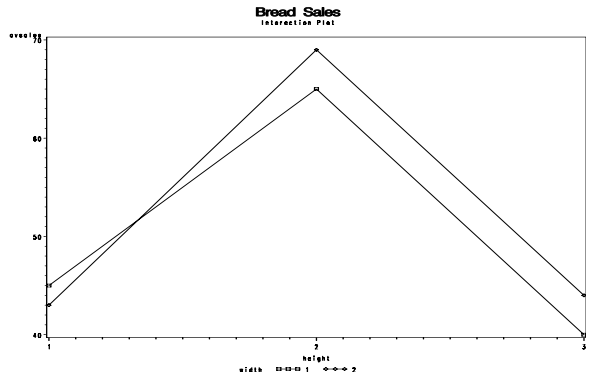
Plot the means

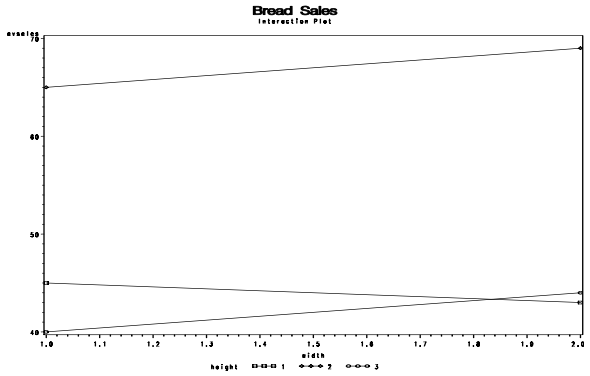
Recall the plotting syntax to get two separate lines for the two width levels. We can also do a plot of sales vs width with three lines for the three heights.

This type of plot is called an "interaction plot" for reasons that we will see later.

```
symbol1 v=square i=join c=black;
symbol2 v=diamond i=join c=black;
symbol3 v=circle i=join c=black;
proc gplot data=avbread;
   plot avsales*height=width;
   plot avsales*width=height;
```

The Interaction plots





Questions

Does the height of the display affect sales? If yes, compare top with middle, top with bottom, and middle with bottom.
Does the width of the display affect sales?
Does the effect of height on sales depend on the width?
Does the effect of width on sales depend on the height?
If yes to the last two, that is an interaction.
Notice that these questions are not straightforward to answer using the cell means model.

Factor Effects Model

For the one-way ANOVA model, we wrote $\mu_i = \mu + \tau_i$ where τ_i was the factor effect. For the two-way ANOVA model, we have $\mu_{i,j} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j}$, where

- μ is the overall (grand) mean - it is $\mu_{..}$ in KNNL
- α_i is the *main effect* of Factor A
- β_j is the *main effect* of Factor B
- $(\alpha\beta)_{i,j}$ is the interaction effect between A and B.

Note that $(\alpha\beta)_{i,j}$ is the name of a parameter all on its own and does not refer to the product of α and β.
Thus the factor effects model is $Y_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \epsilon_{i,j,k}$.
A model without the interaction term, i.e. $\mu_{i,j} = \mu + \alpha_i + \beta_j$, is called an *additive* model.

Parameter Definitions

The overall mean is $\mu = \mu_{..} = \frac{\sum_{i,j}\mu_{i,j}}{ab}$ under the zero-sum constraint (or $\mu = \mu_{ab}$ under the "last = 0 constraint").

The mean for the ith level of A is $\mu_{i.} = \frac{\sum_j \mu_{i,j}}{b}$, and the mean for the jth level of B is $\mu_{.j} = \frac{\sum_i \mu_{i,j}}{a}$.

$\alpha_i = \mu_{i.} - \mu$ and $\beta_j = \mu_{.j} - \mu$, so $\mu_{i.} = \mu + \alpha_i$ and $\mu_{.j} = \mu + \beta_j$.
Note that the α's and β's act like the τ's in the single-factor ANOVA model.
$(\alpha\beta)_{i,j}$ is the difference between $\mu_{i,j}$ and $\mu + \alpha_i + \beta_j$:

$$\begin{aligned}(\alpha\beta)_{i,j} &= \mu_{i,j} - (\mu + \alpha_i + \beta_j) \\ &= \mu_{i,j} - (\mu + (\mu_{i.} - \mu) + (\mu_{.j} - \mu)) \\ &= \mu_{i,j} - \mu_{i.} - \mu_{.j} + \mu\end{aligned}$$

These equations also spell out the relationship between the cell means $\mu_{i,j}$ and the factor effects model parameters.

Interpretation

$\mu_{i,j} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j}$

- μ is the overall mean
- α_i is an adjustment for level i of A.
- β_j is an adjustment for level j of B.
- $(\alpha\beta)_{i,j}$ is an additional adjustment that takes into account both i and j.

Zero-sum Constraints

As in the one-way model, we now have too many parameters and need now several constraints:

$$\begin{aligned}\alpha_{\cdot} &= \sum_i \alpha_i = 0 \\ \beta_{\cdot} &= \sum_j \beta_j = 0 \\ (\alpha\beta)_{\cdot j} &= \sum_i (\alpha\beta)_{i,j} = 0 \ \forall j \text{ (for all } j) \\ (\alpha\beta)_{i\cdot} &= \sum_j (\alpha\beta)_{i,j} = 0 \ \forall i \text{ (for all } i)\end{aligned}$$

Estimates for Factor-effects model

$$\begin{aligned} \hat{\mu} = \quad & \bar{Y}_{...} \quad && = \frac{\sum_{i,j,k} Y_{i,j,k}}{abn} \\ \hat{\mu}_{i.} = \bar{Y}_{i..} \quad & \text{and} \quad && \hat{\mu}_{.j} = \bar{Y}_{.j.} \\ \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...} \quad & \text{and} \quad && \hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...} \\ & (\hat{\alpha\beta})_{i,j} = \quad && \bar{Y}_{i,j.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...} \end{aligned}$$

SS for ANOVA Table

$$\begin{aligned} SSA = \quad & \textstyle\sum_{i,j,k} \hat{\alpha}_i^2 = \sum_{i,j,k} (\bar{Y}_{i..} - \bar{Y}_{...})^2 = nb \sum_i (\bar{Y}_{i..} - \bar{Y})^2 && \text{factor } A \text{ sum of squares} \\ SSB = \quad & \textstyle\sum_{i,j,k} \hat{\beta}_j^2 = \sum_{i,j,k} (\bar{Y}_{.j.} - \bar{Y}_{...})^2 = na \sum_j (\bar{Y}_{.j.} - \bar{Y})^2 && \text{factor } B \text{ sum of squares} \\ SSAB = \quad & \textstyle\sum_{i,j,k} (\hat{\alpha\beta})_{i,j}^2 = n \sum_{i,j} (\hat{\alpha\beta})_{i,j}^2 && AB \text{ interaction sum of squares} \\ SSE = \quad & \textstyle\sum_{i,j,k} (Y_{i,j,k} - \bar{Y}_{i,j.})^2 = \sum_{i,j,k} e_{i,j,k}^2 && \text{error sum of squares} \\ SST = \quad & \textstyle\sum_{i,j,k} (Y_{i,j,k} - \bar{Y}_{...})^2 && \text{total sum of squares} \\ SSM = \quad & SSA + SSB + SSAB = SST - SSE && \text{model sum of squares} \\ SST = \quad & SSA + SSB + SSAB + SSE = SSM + SSE \end{aligned}$$

df for ANOVA Table

$$\begin{aligned} df_A &= a - 1 \\ df_B &= b - 1 \\ df_{AB} &= (a-1)(b-1) \\ df_E &= ab(n-1) \\ df_T &= abn - 1 = n_T - 1 \\ df_M &= a - 1 + b - 1 + (a-1)(b-1) = ab - 1 \end{aligned}$$

MS for ANOVA Table

(no surprises)

$$\begin{aligned} MSA &= SSA/df_A \\ MSB &= SSB/df_B \\ MSAB &= SSAB/df_{AB} \\ MSE &= SSE/df_E \\ MST &= SST/df_T \\ MSM &= SSM/df_M \end{aligned}$$

Hypotheses for two-way ANOVA

Test for Factor A Effect

$$\begin{aligned} \mathrm{H}_0: &\quad \alpha_i = 0 \quad \text{for all } i \\ \mathrm{H}_a: &\quad \alpha_i \neq 0 \quad \text{for at least one } i \end{aligned}$$

The F statistic for this test is $F_A = MSA/MSE$ and under the null hypothesis this follows an F distribution with df_A, df_E.

Test for Factor B Effect

$$\begin{aligned} \mathrm{H}_0: &\quad \beta_j = 0 \quad \text{for all } j \\ \mathrm{H}_a: &\quad \beta_j \neq 0 \quad \text{for at least one } j \end{aligned}$$

The F statistic for this test is $F_B = MSB/MSE$ and under the null hypothesis this follows an F distribution with df_B, df_E.

Test for Interaction Effect

$$\begin{aligned} \mathrm{H}_0: &\quad (\alpha\beta)_{i,j} = 0 \quad \text{for all } (i,j) \\ \mathrm{H}_a: &\quad (\alpha\beta)_{i,j} \neq 0 \quad \text{for at least one } (i,j) \end{aligned}$$

The F statistic for this test is $F_{AB} = MSAB/MSE$ and under the null hypothesis this follows an F distribution with df_{AB}, df_E.

F-statistics for the tests

Notice that the denominator is always MSE and the denominator df is always df_E; the numerators change depending on the test. This is true as long as the effects are *fixed*. That is to say that the levels of our variables are of intrinsic interest in themselves - they are fixed by the experimenter and not considered to be a sample from a larger population of factor levels. For random effects we would need to do something different (more later).

p-values

- p-values are calculated using the $F_{dfNumerator,dfDenominator}$ distributions.
- If $p \leq 0.05$ we conclude that the effect being tested is statistically significant.

ANOVA Table

`proc glm` gives the summary ANOVA table first (model, error, total), then breaks down the model into its components A, B, and AB.

Source	df	SS	MS	F
Model	$ab-1$	SSM	MSM	MSM/MSE
Error	$ab(n-1)$	SSE	MSE	
Total	$abn-1$	$SSTO$	MST	
A	$a-1$	SSA	MSA	MSA/MSE
B	$b-1$	SSB	MSB	MSB/MSE
AB	$(a-1)(b-1)$	$SSAB$	$MSAB$	$MSAB/MSE$

KNNL Example: ANOVA with `GLM`

```
proc glm data=bread;
   class height width;
   model sales=height width height*width;
```

```
The GLM Procedure
Dependent Variable: sales
                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        5     1580.000000      316.000000      30.58    0.0003
Error                        6       62.000000       10.333333
Corrected Total             11     1642.000000

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
height                       2     1544.000000      772.000000      74.71    <.0001
width                        1       12.000000       12.000000       1.16    0.3226
height*width                 2       24.000000       12.000000       1.16    0.3747

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
height                       2     1544.000000      772.000000      74.71    <.0001
width                        1       12.000000       12.000000       1.16    0.3226
height*width                 2       24.000000       12.000000       1.16    0.3747
```

Sums of Squares

- Type I SS are again the sequential sums of squares (variables added in order). Thus height explains 1544, width explains 12 of what is left, and the interaction explains 24 of what is left after that.
- Type III SS is like Type II SS (variable added last) but it also adjusts for differing $n_{i,j}$. So if all cells have the same number of observations (balanced designs are nice - the variables height and width in our example are independent - no multicollinearity!) $SS1$, $SS2$, and $SS3$ will all be the same.
- More details on SS later.

Other output

```
R-Square     Coeff Var      Root MSE    sales Mean
0.962241      6.303040      3.214550      51.00000
```

Results

- The interaction between height and width is not statistically significant ($F = 1.16$; $df = (2, 6)$; $p = 0.37$). *NOTE: Check Interaction FIRST! If it is significant then main effects are left in the model, even if not significant themselves!* We may now go on to examine main effects since our interaction is not significant.
- The main effect of height is statistically significant ($F = 74.71$; $df = (2, 6)$; $p = 4.75 \times 10^{-5}$).
- The main effect of width is not statistically significant ($F = 1.16$; $df = (1, 6)$; $p = 0.32$)

Interpretation

- The height of the display affects sales of bread.
- The width of the display has no apparent effect.
- The effect of the height of the display is similar for both the regular and the wide widths.

Additional Analyses

- We will need to do additional analyses to understand the height effect (factor A).
- There were three levels: bottom, middle and top. Based on the interaction picture, it appears the middle shelf increases sales.
- We could rerun the data with a one-way anova and use the methods we learned in the previous chapters to show this (e.g. tukey)..

Parameter Estimation

Cell Means Model

$Y_{i,j,k} = \mu_{i,j} + \epsilon_{i,j,k}$, where

- $\mu_{i,j}$ is the theoretical mean or expected value of all observations in cell (i, j).
- $\epsilon_{i,j,k} \sim^{iid} N(0, \sigma^2)$
- $Y_{i,j,k} \sim N(\mu_{i,j,k}, \sigma^2)$ are independent

- There are $ab + 1$ parameters of the model: $\mu_{i,j}$; $i = 1, \ldots, a$, $j = 1, \ldots, b$ and σ^2.

For the bread example, estimate the $\mu_{i,j}$ with $\bar{Y}_{i,j.}$ which we can get from the `means height*width` statement:

$$\begin{aligned}
\hat{\mu}_{1,1} &= \bar{Y}_{1,1.} = 45 \\
\hat{\mu}_{1,2} &= \bar{Y}_{1,2.} = 43 \\
\hat{\mu}_{2,1} &= \bar{Y}_{2,1.} = 65 \\
\hat{\mu}_{2,2} &= \bar{Y}_{2,2.} = 69 \\
\hat{\mu}_{3,1} &= \bar{Y}_{3,1.} = 40 \\
\hat{\mu}_{3,2} &= \bar{Y}_{3,2.} = 44
\end{aligned}$$

As usual, σ^2 is estimated by MSE.

Factor Effects Model

$\mu_{i,j} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j}$, where

- μ is the overall (grand) mean - it is $\mu_{..}$ in KNNL
- α_i is the main effect of Factor A
- β_j is the main effect of Factor B
- $(\alpha\beta)_{i,j}$ is the interaction effect between A and B. Note that $(\alpha\beta)_{i,j}$ is the name of a parameter all on its own and does not refer to the product of α and β.

Overall Mean

The overall mean is estimated as $\hat{\mu} = \bar{Y}_{...} = 51$ under the zero-sum constraining. (This is sales mean in the `glm` output). You can get a whole dataset of this value by using a `model` statement with no right hand side, e.g. `model sales=`; and storing the predicted values.

Main Effects

The main effect of A is estimated from the `means` height output. You can get a whole dataset of the $\hat{\mu}_{i.}$ by running a model with just A, e.g. `model sales = height`; and storing the predicted values.
To estimate the α's, you then subtract $\hat{\mu}$ from each height mean.

$$\begin{aligned}
\hat{\mu}_{1.} = \bar{Y}_{1..} = 44 &\Rightarrow \hat{\alpha}_1 = 44 - 51 = -7 \\
\hat{\mu}_{2.} = \bar{Y}_{2..} = 67 &\Rightarrow \hat{\alpha}_2 = 67 - 51 = +16 \\
\hat{\mu}_{3.} = \bar{Y}_{3..} = 42 &\Rightarrow \hat{\alpha}_3 = 42 - 51 = -9
\end{aligned}$$

This says that "middle" shelf height has the effect of a relative increase in sales by 16, while bottom and top decrease the sales by 7 and 9 respectively. Notice that these sum to zero so that there is no "net" effect (there are only 2 free parameters, $df_A = 2$).

The main effect of B is similarly estimated from the `means width` output, or by storing the predicted values of `model sales = width`; then subtract $\hat{\mu}$ from each height mean.

$$\begin{aligned} \hat{\mu}_{.1} = \bar{Y}_{.1.} = 50 \quad &\Rightarrow \quad \hat{\beta}_1 = 50 - 51 = -1 \\ \hat{\mu}_{.2} = \bar{Y}_{.2.} = 52 \quad &\Rightarrow \quad \hat{\beta}_2 = 52 - 51 = +1 \end{aligned}$$

Wide display increases sales by an average of 1, while regular display decreases sales by 1 (they sum to zero so there's only 1 free parameter, $df_B = 1$).

Interaction Effects

Recall that $\hat{\alpha\beta}_{i,j} = \hat{\mu}_{i,j} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)$. This is the difference between the treatment mean and the value predicted by the overall mean and main effects only (i.e. by the additive model). You can get the treatment means from the `means height*width` statement, or by the predicted values of `model sales=height*width`; then subtract the appropriate combination of the previously estimated parameters.

$$\begin{aligned} (\hat{\alpha\beta})_{11} &= \bar{Y}_{11.} - (\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1) = 45 - (51 - 7 - 1) = +2 \\ (\hat{\alpha\beta})_{12} &= \bar{Y}_{12.} - (\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_2) = 43 - (51 - 7 + 1) = -2 \\ (\hat{\alpha\beta})_{21} &= \bar{Y}_{21.} - (\hat{\mu} + \hat{\alpha}_2 + \hat{\beta}_1) = 65 - (51 + 16 - 1) = -1 \\ (\hat{\alpha\beta})_{22} &= \bar{Y}_{22.} - (\hat{\mu} + \hat{\alpha}_2 + \hat{\beta}_2) = 69 - (51 + 16 + 1) = +1 \\ (\hat{\alpha\beta})_{31} &= \bar{Y}_{31.} - (\hat{\mu} + \hat{\alpha}_3 + \hat{\beta}_1) = 40 - (51 - 9 - 1) = -1 \\ (\hat{\alpha\beta})_{32} &= \bar{Y}_{32.} - (\hat{\mu} + \hat{\alpha}_3 + \hat{\beta}_2) = 44 - (51 - 9 + 1) = +1 \end{aligned}$$

Notice that they sum in pairs (over j) to zero and also the sum over i is zero for each j. Thus there are in reality only two free parameters here ($df_{AB} = 2$).

Doing this in SAS

Unlike `proc reg`, you are only allowed one `model` statement per call to `glm`. But this at least saves you doing the arithmetic by hand.

```
proc glm data=bread;
        class height width;
        model sales=;
        output out=pmu p=muhat;
proc glm data=bread;
        class height width;
```

```
        model sales=height;
        output out=pA p=Amean;
proc glm data=bread;
        class height width;
        model sales=width;
        output out=pB p=Bmean;
proc glm data=bread;
        class height width;
        model sales=height*width;
        output out=pAB p=ABmean;
data parmest;
        merge bread pmu pA pB pAB;
        alpha = Amean - muhat;
        beta = Bmean - muhat;
        alphabeta = ABmean - (muhat+alpha+beta);
proc print data=parmest;
```

Obs	sales	height	width	muhat	Amean	Bmean	ABmean	alpha	beta	alphabeta
1	47	1	1	51	44	50	45	-7	-1	2
2	43	1	1	51	44	50	45	-7	-1	2
3	46	1	2	51	44	52	43	-7	1	-2
4	40	1	2	51	44	52	43	-7	1	-2
5	62	2	1	51	67	50	65	16	-1	-1
6	68	2	1	51	67	50	65	16	-1	-1
7	67	2	2	51	67	52	69	16	1	1
8	71	2	2	51	67	52	69	16	1	1
9	41	3	1	51	42	50	40	-9	-1	-1
10	39	3	1	51	42	50	40	-9	-1	-1
11	42	3	2	51	42	52	44	-9	1	1
11	46	3	2	51	42	52	44	-9	1	1

Zero-sum Constraints

$$\alpha_\cdot = \sum_i \alpha_i = 0$$

$$\beta_\cdot = \sum_j \beta_j = 0$$

$$(\alpha\beta)_{\cdot j} = \sum_i (\alpha\beta)_{i,j} = 0 \quad \forall j \quad \text{(for all } j\text{)}$$

$$(\alpha\beta)_{\cdot i} = \sum_i (\alpha\beta)_{i,j} = 0 \quad \forall i \quad \text{(for all } i\text{)}$$

All of these constraints are satisfied by the above estimates.

Notice how these main and interaction effects fit together to give back the treatment means:

$$\begin{aligned} 45 &= 51 - 7 - 1 + 2 \\ 43 &= 51 - 7 + 1 - 2 \\ 65 &= 51 + 16 - 1 - 1 \\ 69 &= 51 + 16 + 1 + 1 \\ 40 &= 51 - 9 - 1 - 1 \\ 44 &= 51 - 9 + 1 + 1 \end{aligned}$$

SAS `GLM` Constraints

As usual, SAS has to do its constraints differently. As in one-way ANOVA, it sets the parameter for the last category equal to zero.

$$\begin{aligned} \alpha_a = 0 \quad & (1 \text{ constraint}) \\ \beta_b = 0 \quad & (1 \text{ constraint}) \\ (\alpha\beta)_{a,j} = 0 \quad & \text{for all } j \qquad (b \text{ constraints}) \\ (\alpha\beta)_{i,b} = 0 \quad & \text{for all } i \qquad (a \text{ constraints}) \end{aligned}$$

The total is $1 + 1 + a + b - 1 = a + b + 1$ constraints (the constraint $(\alpha\beta)_{a,b}$ is counted twice above).

Parameters and constraints

The cell means model has ab parameters for the means. The factor effects model has $(1 + a + b + ab)$ parameters.

- An intercept (1)
- Main effect of A (a)
- Main effect of B (b)
- Interaction of A and B (ab)

There are $1 + a + b + ab$ parameters and $1 + a + b$ constraints, so there are ab remaining unconstrained parameters (or sets of parameters), the same number of parameters for the means in the cell means model. This is the number of parameters we can actually estimate.

KNNL Example

KNNL page 823 (`nknw817b.sas`)
Y is the number of cases of bread sold
A is the height of the shelf display, $a = 3$ levels: bottom, middle, top
B is the width of the shelf display, $b = 2$: regular, wide
$n = 2$ stores for each of the 3×2 treatment combinations

`proc glm` with solution

We will get *different estimates* for the parameters here because a different constraint system is used.

```
proc glm data=bread;
   class height width;
   model sales=height width height*width/solution;
   means height*width;
```

`Solution` output

Intercept	44.00000000 B*	=	$\hat{\mu}$
height 1	-1.00000000 B	=	$\hat{\alpha}_1$
height 2	25.00000000 B*	=	$\hat{\alpha}_2$
height 3	0.00000000 B	=	$\hat{\alpha}_3$
width 1	-4.00000000 B	=	$\hat{\beta}_1$
width 2	0.00000000 B	=	$\hat{\beta}_2$
height*width 1 1	6.00000000 B	=	$(\hat{\alpha\beta})_{1,1}$
height*width 1 2	0.00000000 B	=	$(\hat{\alpha\beta})_{1,2}$
height*width 2 1	0.00000000 B	=	$(\hat{\alpha\beta})_{2,1}$
height*width 2 2	0.00000000 B	=	$(\hat{\alpha\beta})_{2,2}$
height*width 3 1	0.00000000 B	=	$(\hat{\alpha\beta})_{3,1}$
height*width 3 2	0.00000000 B	=	$(\hat{\alpha\beta})_{3,2}$

It also prints out standard errors, t-tests and p-values for testing whether each parameter is equal to zero. That output has been omitted here but the significant ones have been starred. Notice that the last α and β are set to zero, as well as the last $\hat{\alpha\beta}$ in each category. *They no longer sum to zero.*

Means

The estimated treatment means are $\hat{\mu}_{i,j} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + (\hat{\alpha\beta})_{i,j}$.

```
height      width       N            Mean
1           1           2      45.0000000 = 44 - 1 - 4 + 6
```

```
1           2           2       43.0000000 = 44 - 1 + 0 + 0
2           1           2       65.0000000 = 44 + 25 - 4 + 0
2           2           2       69.0000000 = 44 + 25 + 0 + 0
3           1           2       40.0000000 = 44 + 0 - 4 + 0
3           2           2       44.0000000 = 44 + 0 + 0 + 0
```

ANOVA Table

Source	df	SS	MS	F
A	$a-1$	SSA	MSA	MSA/MSE
B	$b-1$	SSB	MSB	MSB/MSE
AB	$(a-1)(b-1)$	$SSAB$	$MSAB$	$MSAB/MSE$
Error	$ab(n-1)$	SSE	MSE	
Total	$abn-1$	$SSTO$	MST	

Expected Mean Squares

$$\begin{aligned}
\mathrm{E}(MSE) &= \sigma^2 \\
\mathrm{E}(MSA) &= \sigma^2 + \frac{nb}{a-1}\sum_i \alpha_i^2 \\
\mathrm{E}(MSB) &= \sigma^2 + \frac{na}{b-1}\sum_j \beta_j^2 \\
\mathrm{E}(MSAB) &= \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i,j}(\alpha\beta)_{i,j}^2
\end{aligned}$$

Here, α_i, β_j, and $(\alpha\beta)_{i,j}$ are defined with the usual zero-sum constraints.

Analytical strategies

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

What if AB interaction is not significant?

If the AB interaction is not statistically significant, you could rerun the analysis without the interaction (see discussion of pooling KNNL Section 19.10). This will put the SS and *df* for AB into Error. Results of main effect hypothesis tests could change because MSE and denominator *df* have changed (more impact with small sample size). If one main effect 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.

If neither main effect is significant...

- Model could be run as `Y=`; (i.e. no factors at all)
- A one population model
- This seems silly, but this syntax can be useful for getting parameter estimates in the preferred constraint system (see below).

For a main effect with more than two levels that is significant, use the `means` statement with the Tukey multiple comparison procedure. Contrasts and linear combinations can also be examined using the `contrast` and `estimate` statements.

If AB interaction is significant 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.
- For example, the interaction effect may be much smaller in magnitude than the main effects; or may only be apparent in a small number of treatments.
- Use the marginal means for each significant main effect to describe the important results for the main effects.
- You may need to qualify these results using the interaction.
- Keep the interaction in the model.
- Carefully interpret the marginal means as averages over the levels of the other factor.
- KNNL also discuss ways that transformations can sometimes eliminate interactions.

If AB interaction is significant and important

The interaction effect is so large and/or pervasive that main effects cannot be interpreted on their own.
Options include the following:

- Treat as a one-way ANOVA with ab levels; use Tukey to compare means; contrasts and estimate can also be useful.
- Report that the interaction is significant; plot the means and describe the pattern.
- Analyze the levels of A for each level of B (use a `by` statement) or vice versa

Strategy for the bread example

Previous results with interaction

```
The GLM Procedure
Dependent Variable: sales
                            Sum of
Source           DF         Squares  Mean Square    F Value    Pr > F
Model             5     1580.000000   316.000000      30.58    0.0003
Error             6       62.000000    10.333333
Corrected Total  11     1642.000000

Source        DF       Type I SS     Mean Square    F Value    Pr > F
height         2     1544.000000      772.000000      74.71    <.0001
width          1       12.000000       12.000000       1.16    0.3226
height*width   2       24.000000       12.000000       1.16    0.3747
```

Rerun without interaction

```
proc glm data=bread;
   class height width;
   model sales=height width;
   means height / tukey lines;
```

```
                                    Sum of
Source                       DF         Squares     Mean Square    F Value    Pr > F
Model                         3     1556.000000      518.666667      48.25    <.0001
Error                         8       86.000000       10.750000
Corrected Total              11     1642.000000

Source                       DF       Type I SS     Mean Square    F Value    Pr > F
height                        2     1544.000000      772.000000      71.81    <.0001
width                         1       12.000000       12.000000       1.12    0.3216
```

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. A drawback is that if the SS for that term is large it will increase the SSE too much. Therefore we would only want to do this for insignificant terms, i.e. those with small SS, most often the interaction term.
This strategy can be beneficial in small experiments where the df_E is very small.
Do not remove the main effect and leave the interaction term. Typically we are pooling SSE and $SSAB$.

In our example, MSh and MSw have not changed, but MSE, F's, and p-values have

changed. In this case MSE went up, but in other cases it might go down.
Note SSE: $62 + 24 = 86$; and df_E: $6 + 2 = 8$.

Tukey Output

```
           Mean      N    height

A        67.000      4    2

B        44.000      4    1
B
B        42.000      4    3
```

As we noticed from the plot, the middle shelf is significantly different (better in terms of sales if we look at the plot) from the other two.

Specification of `contrast` and `estimate` statements

- When using the `contrast` statement, you can double check your results with the `estimate` statement.
- The order of factors is determined by the order in the `class` statement, not the order in the `model` statement.
- **Contrasts to be examined should come *a priori* from research questions, not from questions that arise after looking at the plots and means.**

`Contrast/Estimate` Example

For the bread example, suppose we want to compare the average of the two height = middle cells with the average of the other four cells; i.e. look at "eye-level" vs. "not eye-level" (for the average person). With this approach, the contrast should correspond to a research question formulated before examining the data. First, formulate the question as a contrast in terms of the cell means model:

$$\mathrm{H}_0 : \frac{(\mu_{2,1} + \mu_{2,2})}{2} = \frac{(\mu_{1,1} + \mu_{1,2} + \mu_{3,1} + \mu_{3,2})}{4}$$

or $L = 0$, where

$$L = -0.25\mu_{1,1} - 0.25\mu_{1,2} + 0.5\mu_{2,1} + 0.5\mu_{2,2} - 0.25\mu_{3,1} - 0.25\mu_{3,2}$$

Then translate the contrast into the factor effects model using $\mu_{i,j} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j}$

$$\begin{aligned}
-0.25\mu_{1,1} &= -0.25(\mu + \alpha_1 + \beta_1 + \alpha\beta_{1,1}) \\
-0.25\mu_{1,2} &= -0.25(\mu + \alpha_1 + \beta_2 + \alpha\beta_{1,2}) \\
0.50\mu_{2,1} &= +0.50(\mu + \alpha_2 + \beta_1 + \alpha\beta_{2,1}) \\
0.50\mu_{2,2} &= +0.50(\mu + \alpha_2 + \beta_2 + \alpha\beta_{2,2}) \\
-0.25\mu_{3,1} &= -0.25(\mu + \alpha_3 + \beta_1 + \alpha\beta_{3,1}) \\
-0.25\mu_{3,2} &= -0.25(\mu + \alpha_3 + \beta_2 + \alpha\beta_{3,2})
\end{aligned}$$

$$\begin{aligned}
L = &(-0.5\alpha_1 + \alpha_2 - 0.5\alpha_3) \\
&+(-0.25\alpha\beta_{1,1} - 0.25\alpha\beta_{1,2} + 0.5\alpha\beta_{2,1} + 0.5\alpha\beta_{2,2} - 0.25\alpha\beta_{3,1} - 0.25\alpha\beta_{3,2})
\end{aligned}$$

Note the β's do not appear in this contrast because we are looking at height only and averaging over width (this would not necessarily be true in an unbalanced design).

`proc glm` with `contrast` and `estimate`

(`nknw864.sas`)

```
proc glm data=bread;
   class height width;
   model sales=height width height*width;
   contrast 'middle vs others'
      height -.5 1 -.5
      height*width -.25 -.25 .5 .5 -.25 -.25;
   estimate 'middle vs others'
      height -.5 1 -.5
      height*width -.25 -.25 .5 .5 -.25 -.25;
   means height*width;
```

Output

```
Contrast                    DF     Contrast SS     Mean Square    F Value    Pr > F
middle vs others             1     1536.000000     1536.000000     148.65    <.0001


                                                  Standard
Parameter                    Estimate                Error    t Value    Pr > |t|
middle vs others          24.0000000          1.96850197      12.19      <.0001
```

Check with `means`

```
1  1  45
1  2  43
2  1  65
2  2  69
3  1  40
3  2  44
```

$$\hat{L} = \frac{(65+69)}{2} - \frac{(45+43+40+44)}{4} = 24$$

Combining with Quantitative Factors

Sometimes a factor can be interpreted as either categorical or quantitative. For example, "low, medium, high" or actual height above floor. If there are replicates for a quantitative factor we could use either regression or ANOVA. Recall that `GLM` will treat a factor as quantitative unless it is listed in the `class` statement. Notice that you can use ANOVA even if the relationship with the quantitative variable is non-linear, whereas with regression you would have to find that relationship.

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; i.e, interaction terms.
- Lack of fit analysis can be useful (recall `trainhrs` example).

Two Quantitative factors

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

Chapter 20: One Observation per Cell

For $Y_{i,j,k}$, as usual

- i denotes the level of the factor A
- j denotes the level of the factor B
- k denotes the kth observation in cell (i, j)
- $i = 1, \ldots, a$ levels of factor A

- $j = 1, \ldots, b$ levels of factor B

Now suppose we have $n = 1$ observation in each cell (i, j). We can no longer estimate variances separately for each treatment. *The impact is that we will not be able to estimate the interaction terms; we will have to assume no interaction.*

Factor Effects Model

$\mu_{i,j} = \mu + \alpha_i + \beta_j$

- 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. We assume no interaction.

Constraints

- Text: $\sum \alpha_i = 0$ and $\sum \beta_j = 0$
- SAS `glm`: $\alpha_a = \beta_b = 0$

ANOVA Table

Source	df	SS	MS	F
A	$a-1$	SSA	MSA	MSA/MSE
B	$b-1$	SSB	MSB	MSB/MSE
Error	$(a-1)(b-1)$	SSE	MSE	
Total	$ab-1$	$SSTO$	MST	

Expected Mean Squares

$$\begin{aligned} \mathrm{E}(MSE) &= \sigma^2 \\ \mathrm{E}(MSA) &= \sigma^2 + \frac{b}{a-1}\sum_i \alpha_i^2 \\ \mathrm{E}(MSB) &= \sigma^2 + \frac{a}{b-1}\sum_j \beta_j^2 \end{aligned}$$

Here, α_i and β_j are defined with the zero-sum factor effects constraints.

KNNL Example

- KNNL page 882 (`nknw878.sas`)
- Y is the premium for auto insurance
- A is the size of the city, $a = 3$ levels: small, medium and large
- B is the region, $b = 2$: East, West
- $n = 1$
- the response is the premium charged by a particular company

The data

```
data carins;
infile 'H:\System\Desktop\CH21TA02.DAT';
   input premium size region;
if size=1 then sizea='1_small ';
if size=2 then sizea='2_medium';
if size=3 then sizea='3_large ';
proc print data=carins;
```

```
Obs    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
```

```
proc glm data=carins;
   class sizea region;
   model premium=sizea region/solution;
   means sizea region / tukey;
   output out=preds p=muhat;
```

```
The GLM Procedure

             Class Level Information
Class         Levels    Values
sizea              3    1_small 2_medium 3_large
region             2    1 2

Number of observations    6

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        3     10650.00000      3550.00000      71.00    0.0139
```

Error	2	100.00000	50.00000
Corrected Total	5	10750.00000	

Notice that we only have 5 total df. If we had interaction in the model it would use up another 2 df and there would be 0 left to estimate error.

R-Square	Coeff Var	Root MSE	premium Mean
0.990698	4.040610	7.071068	175.0000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sizea	2	9300.000000	4650.000000	93.00	0.0106
region	1	1350.000000	1350.000000	27.00	0.0351

Both main effects are significant.

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		195.0000000	B	5.77350269	33.77	0.0009
sizea	1_small	-90.0000000	B	7.07106781	-12.73	0.0061
sizea	2_medium	-15.0000000	B	7.07106781	-2.12	0.1679
sizea	3_large	0.0000000	B	.	.	.
region	1	30.0000000	B	5.77350269	5.20	0.0351
region	2	0.0000000	B	.	.	.

Check vs predicted values ($\hat{\mu}$)

region	sizea	muhat
1	1_small	$135 = 195 - 90 + 30$
2	1_small	$105 = 195 - 90$
1	2_medium	$210 = 195 - 15 + 30$
2	2_medium	$180 = 195 - 15$
1	3_large	$225 = 195 + 30$
2	3_large	$195 = 195$

Multiple Comparisons Size

	Mean	N	sizea
A	210.000	2	3_large
A			
A	195.000	2	2_medium
B	120.000	2	1_small

The ANOVA results told us that size was significant; now we additionally know that small is different from medium and large, but that medium and large do not differ significantly.

Multiple Comparisons Region

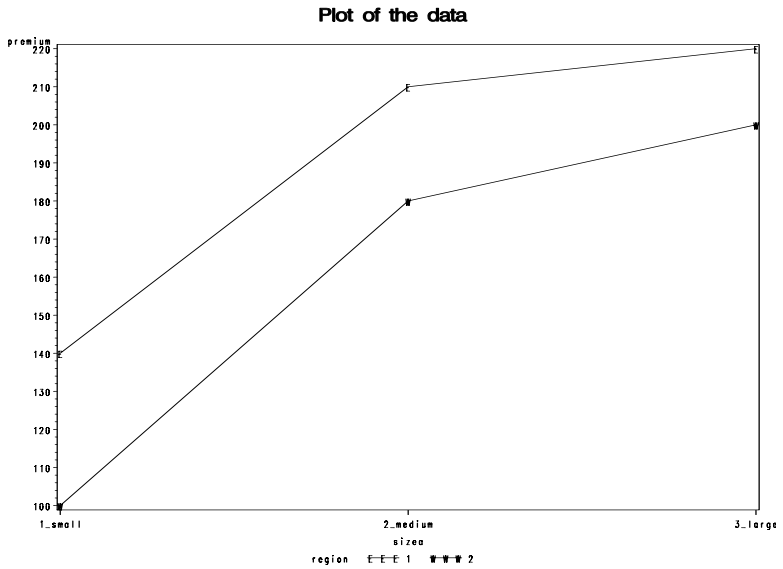
	Mean	N	region
A	190.000	3	1
B	160.000	3	2

The ANOVA results told us that these were different since region was significant (only two levels) ...

So this gives us no new information.

Plot the data

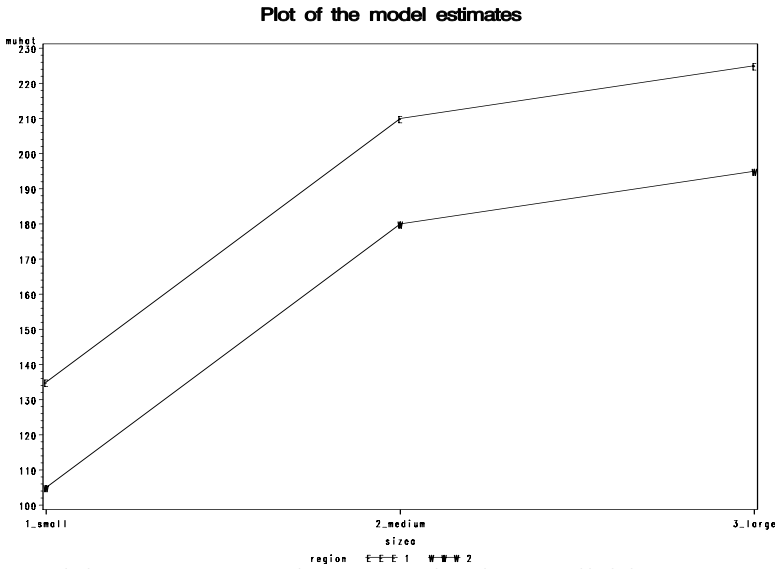
```
symbol1 v='E' i=join c=black;
symbol2 v='W' i=join c=black;
title1 'Plot of the data';
proc gplot data=preds;
   plot premium*sizea=region;
```



The lines are not quite parallel, but the interaction, if any, does not appear to be substantial. If it was, our analysis would not be valid and we would need to collect more data.

Plot the estimated model

```
title1 'Plot of the model estimates';
proc gplot data=preds;
   plot muhat*sizea=region;
```



Notice that the model estimates produce completely parallel lines.

Tukey test for additivity

If we believe interaction is a problem, this is a possible way to test it without using up all our df.
One additional term is added to the model (θ), replacing the $(\alpha\beta)_{i,j}$ with the product:

$$\mu_{i,j} = \mu + \alpha_i + \beta_j + \theta\alpha_i\beta_j$$

We use one degree of freedom to estimate θ, leaving one left to estimate error. Of course, this only tests for interaction of the specified form, but it may be better than nothing.

There are other variations on this idea, such as $\theta_i\beta_j$.

Find $\hat{\mu}$ (grand mean)

(`nknw884.sas`)

```
proc glm data=carins;
   model premium=;
   output out=overall p=muhat;
proc print data=overall;
```

Obs	premium	size	region	muhat
1	140	1	1	175
2	100	1	2	175
3	210	2	1	175
4	180	2	2	175
5	220	3	1	175
6	200	3	2	175

Find $\hat{\mu}_A$ (treatment means)

```
proc glm data=carins;
   class size;
   model premium=size;
   output out=meanA p=muhatA;
proc print data=meanA;
```

Obs	premium	size	region	muhat A
1	140	1	1	120
2	100	1	2	120
3	210	2	1	195
4	180	2	2	195
5	220	3	1	210
6	200	3	2	210

Find $\hat{\mu}_B$ (treatment means)

```
proc glm data=carins;
   class region;
   model premium=region;
   output out=meanB p=muhatB;
proc print data=meanB;
```

Obs	premium	size	region	muhat B
1	140	1	1	190
2	100	1	2	160
3	210	2	1	190
4	180	2	2	160
5	220	3	1	190
6	200	3	2	160

Combine and Compute

```
data estimates;
   merge overall meanA meanB;
   alpha = muhatA - muhat;
   beta = muhatB - muhat;
   atimesb = alpha*beta;
proc print data=estimates;
   var size region alpha beta atimesb;
```

Obs	size	region	alpha	beta	atimesb
1	1	1	-55	15	-825
2	1	2	-55	-15	825

```
 3       2        1        20       15       300
 4       2        2        20      -15      -300
 5       3        1        35       15       525
 6       3        2        35      -15      -525

proc glm data=estimates;
   class size region;
   model premium=size region atimesb/solution;

                                      Sum of
Source                       DF         Squares     Mean Square    F Value    Pr > F
Model                         4     10737.09677      2684.27419     208.03    0.0519
Error                         1        12.90323        12.90323
Corrected Total               5     10750.00000

R-Square     Coeff Var      Root MSE    premium Mean
0.998800      2.052632      3.592106        175.0000

Source                       DF       Type I SS     Mean Square    F Value    Pr > F
size                          2     9300.000000     4650.000000     360.37    0.0372
region                        1     1350.000000     1350.000000     104.62    0.0620
atimesb                       1       87.096774       87.096774       6.75    0.2339

                                        Standard
Parameter            Estimate              Error    t Value    Pr > |t|
Intercept        195.0000000 B      2.93294230      66.49      0.0096
size      1      -90.0000000 B      3.59210604     -25.05      0.0254
size      2      -15.0000000 B      3.59210604      -4.18      0.1496
size      3        0.0000000 B       .                .         .
region    1       30.0000000 B      2.93294230      10.23      0.0620
region    2        0.0000000 B       .                .         .
atimesb           -0.0064516        0.00248323      -2.60      0.2339
```

The test for `atimesb` is testing $H_0 : \theta = 0$, which is not rejected. According to this, the interaction is not significant. Notice the increased p-values on the main effect tests, because we used up a df.