Multi-way ANOVA

CONCEPTS FROM EXPERIMENTAL DESIGN

TWO-WAY ANOVA

Example: Crop Yields

The Experimental Design

Suppose we have two factors A and B
(Say, A is amount of fertilizer (J = 2) and B is amount of manure (K = 2))

Due to the small number of levels of each factor, we will consider a crossed (factorial) design, where each level of A and B jointly occur

We will measure the total plant matter yield of each plant

Here, the experimental unit will be an individual plant receiving the treatments

After talking with the scientist, we decide that she believes an interaction between A and B is likely.

The scientist thinks 5 replications should be enough

Power Analysis

The next step should always be a power analysis

Hopefully it will determine the set-up of the experiment

But at least it will provide added information if there is no estimated effect

```
DATA Exemplary;
    INPUT fert $ manure $ yield;
    DATALINES;
        High High 16
        High Low 15
        Low High 15
        Low Low 12
    ;
RUN;

PROC GLMPOWER DATA = Exemplary;
    CLASS fert manure;
    MODEL yield = fert manure fert*manure;
    POWER
         STDDEV = 1.5
         NTOTAL = 20
         POWER  = .;
RUN;
```

We need to elicit:

- Effect sizes we wish to detect
- The considered model
- The expected variability of the experiment

The GLMPOWER Procedure

Fixed Scenario Elements	
Dependent Variable	yield
Error Standard Deviation	1.5
Total Sample Size	20
Alpha	0.05
Error Degrees of Freedom	16

Computed Power			
Index	Source	Test DF	Power
1	fert	1	0.799
2	manure	1	0.799
3	fert*manure	1	0.289

Power Analysis

What about this scenario (hint: sketch the profile plot):

```
DATA Exemplary;
     INPUT fert $ manure $ yield;
     DATALINES;
          High High 18
          High Low 15
          Low High 15
          Low Low 12
     ;
RUN;

PROC GLMPOWER DATA = Exemplary;
     CLASS fert manure;
     MODEL yield = fert manure fert*manure;
     POWER
          STDDEV = 1.5
          NTOTAL = 20
          POWER  = .;
RUN;
```

The GLMPOWER Procedure

Fixed Scenario Elements	
Dependent Variable	yield
Error Standard Deviation	1.5
Total Sample Size	20
Alpha	0.05
Error Degrees of Freedom	16

Computed Power			
Index	Source	Test DF	Power
1	fert	1	0.987
2	manure	1	0.987
3	fert*manure	1	0.050

The Experiment

After discussing the results of the power analysis, we agree that 5 replications and a crossed design is appropriate….

For the next few weeks, the experiment is conducted..

Statisticians

Scientists

The Data

After discussing the results of the power analysis, we agree that 5 replications and a crossed design is appropriate….

For the next few weeks, the experiment is conducted..

```
DATA cropYield;
    INPUT Fert $ Manure $ Yield;
    DATALINES;
    High High 13.7
    High High 16.8
    High High 14.9
    High High 17.6
    High High 16.5
    Low High 17.4
    Low High 13.5
    Low High 15.1
    Low High 15.4
    Low High 13.2
    High Low 16.0
    High Low 16.1
    High Low 13.0
    High Low 16.7
    High Low 13.2
    Low Low 13.4
    Low Low 11.6
    Low Low 14.7
    Low Low 9.7
    Low Low 11.9
    ;
RUN;
```

Detour: Notation

The Notation

There are two factors: A and B with two levels each $J = K = 2$

We can equivalently express the interaction model in three major ways

They're all representing the same model: $\mu\{Y|$ A, B$\}$ = A + B + A*B

- "ANOVA representation"

$$Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk}$$

- "Effects representation"

$$Y_{ijk} = \mu_j + \mu_k + \varepsilon_{ijk}$$

- "Regression representation"

$$Y_i = \beta_0 + \beta_1 Fert + \beta_2 Manure + \beta_3 Fert * Manure + \varepsilon_i$$

Important: these are all linear models of the form $\mathbb{Y} = \mathbb{X}\beta + \varepsilon$

All three models impose additional constraints for estimation

Let's explore these representations as they affect the code and design matrix $\mathbb{X}$

ANOVA representation

$Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk}$

The reason the ANOVA model is usually written like this is that it directly encodes the core ANOVA test:

H_0: The equal means model is sufficient ($Y_{ijk} = \mu + \varepsilon_{ijk}$)

H_A: The equal means model isn't sufficient

The draw back of this representation:

It is over-parameterized, which in statistics is referred to as **UNIDENTIFIED**

This means that different settings for the parameters cannot be detected via the data

Let's look at an example

ANOVA representation

$Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk}$

Consider two possible true (though unknown) states of nature:

$\mu = 10$

- $\mu_{j=1} = 3$
- $\mu_{j=2} = 5$
- $\mu_{k=1} = 0$
- $\mu_{k=2} = 0$
- $\mu_{jk} = 0$ (for all j, k)

$\mu = 7$

- $\mu_{j=1} = 6$
- $\mu_{j=2} = 5$
- $\mu_{k=1} = 0$
- $\mu_{k=2} = 0$
- $\mu_{jk} = 0$ (for all j, k)

The mean of Y for A = 1, B = 2 in both cases is 13
even though the parameter values are different
(Note that some important contrasts are unique, such as $\mu + \mu_{j=1}$)

ANOVA representation

UNIDENTIFIED: Columns are linear combinations

Due to SAS's historical relationship with developing ANOVA, the ANOVA representation is the default for PROC GLM

$$Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk}$$

```
PROC GLMMOD DATA = cropYield;
    CLASS fert manure;
    MODEL yield = fert manure fert*manure;
RUN;
```

The GLMMOD Procedure

Parameter Definitions			
		CLASS Variable Values	
Column Number	Name of Associated Effect	Fert	Manure
1	Intercept		
2	Fert	High	
3	Fert	Low	
4	Manure		High
5	Manure		Low
6	Fert*Manure	High	High
7	Fert*Manure	High	Low
8	Fert*Manure	Low	High
9	Fert*Manure	Low	Low

Design Points										
		Column Number								
Observation Number	Yield	1	2	3	4	5	6	7	8	9
1	13.7	1	1	0	1	0	1	0	0	0
2	16.8	1	1	0	1	0	1	0	0	0
3	14.9	1	1	0	1	0	1	0	0	0
4	17.6	1	1	0	1	0	1	0	0	0
5	16.5	1	1	0	1	0	1	0	0	0
6	17.4	1	0	1	1	0	0	0	1	0
7	13.5	1	0	1	1	0	0	0	1	0
8	15.1	1	0	1	1	0	0	0	1	0
9	15.4	1	0	1	1	0	0	0	1	0
10	13.2	1	0	1	1	0	0	0	1	0
11	16.0	1	1	0	0	1	0	1	0	0
12	16.1	1	1	0	0	1	0	1	0	0
13	13.0	1	1	0	0	1	0	1	0	0
14	16.7	1	1	0	0	1	0	1	0	0
15	13.2	1	1	0	0	1	0	1	0	0
16	13.4	1	0	1	0	1	0	0	0	1
17	11.6	1	0	1	0	1	0	0	0	1
18	14.7	1	0	1	0	1	0	0	0	1
19	9.7	1	0	1	0	1	0	0	0	1
20	11.9	1	0	1	0	1	0	0	0	1

$= \mathbb{X}$

Effects representation

$Y_{ijk} = \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk}$

The effects representation omits the overall mean term μ

Under this specification, the model is **IDENTIFIED** for the one-way model, which is why I'm referring to it at all

However, it is **UNIDENTIFIED** for a general multi-way ANOVA

However, for both the ANOVA and Effects important parameters are identified, even though the model is not:

$\mu_{j_1} - \mu_{j_2}$

$\mu_{k_1} - \mu_{k_2}$

Effects representation

UNIDENTIFIED: Columns are linear combinations

We can get SAS to drop the mean term:

$$Y_{ijk} = \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk}$$

(this is more useful in one-way ANOVA)

```
PROC GLMMOD DATA = cropYield;
    CLASS fert manure;
    MODEL yield = fert manure fert*manure / NOINT;
RUN;
```

The GLMMOD Procedure

Parameter Definitions			
		CLASS Variable Values	
Column Number	Name of Associated Effect	Fert	Manure
1	Fert	High	
2	Fert	Low	
3	Manure		High
4	Manure		Low
5	Fert*Manure	High	High
6	Fert*Manure	High	Low
7	Fert*Manure	Low	High
8	Fert*Manure	Low	Low

Design Points									
		Column Number							
Observation Number	Yield	1	2	3	4	5	6	7	8
1	13.7	1	0	1	0	1	0	0	0
2	16.8	1	0	1	0	1	0	0	0
3	14.9	1	0	1	0	1	0	0	0
4	17.6	1	0	1	0	1	0	0	0
5	16.5	1	0	1	0	1	0	0	0
6	17.4	0	1	1	0	0	0	1	0
7	13.5	0	1	1	0	0	0	1	0
8	15.1	0	1	1	0	0	0	1	0
9	15.4	0	1	1	0	0	0	1	0
10	13.2	0	1	1	0	0	0	1	0
11	16.0	1	0	0	1	0	1	0	0
12	16.1	1	0	0	1	0	1	0	0
13	13.0	1	0	0	1	0	1	0	0
14	16.7	1	0	0	1	0	1	0	0
15	13.2	1	0	0	1	0	1	0	0
16	13.4	0	1	0	1	0	0	0	1
17	11.6	0	1	0	1	0	0	0	1
18	14.7	0	1	0	1	0	0	0	1
19	9.7	0	1	0	1	0	0	0	1
20	11.9	0	1	0	1	0	0	0	1

$= \mathbb{X}$

Regression representation

$Y_i = \beta_0 + \beta_1 Fert + \beta_2 Manure + \beta_3 Fert * Manure + \varepsilon_i$

Here, Fert and Manure are indicators of a specific level (say ”high”) and hence “low” is the reference.

The “indicator variable” specification allows identification via forcing some of the terms in the ANOVA representation to zero

ANOVA representation: $Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk}$

Suppose $j = 1$ is for “high fertilizer” and hence $Fert = 1$

Also, $\mathrm{k} = 1$ is for “high manure” and hence $Manure = 1$

Then $\mu + \mu_{j=1} + \mu_{k=1} + \mu_{j=1,k=1}$ is encoded as $\beta_0 + \beta_1 + \beta_2 + \beta_3$

Regression representation

```
PROC GLM DATA = cropYield PLOTS = all;
    CLASS fert manure;
    MODEL yield = fert manure fert*manure / SOLUTION;
RUN;
```

Extra constraints: Force some of the ANOVA model parameters to zero:

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	12.26000000	B	0.77459667	15.83	<.0001
Fert High	2.74000000	B	1.09544512	2.50	0.0236
Fert Low	0.00000000	B	.	.	.
Manure High	2.66000000	B	1.09544512	2.43	0.0273
Manure Low	0.00000000	B	.	.	.
Fert*Manure High High	-1.76000000	B	1.54919334	-1.14	0.2727
Fert*Manure High Low	0.00000000	B	.	.	.
Fert*Manure Low High	0.00000000	B	.	.	.
Fert*Manure Low Low	0.00000000	B	.	.	.

Back to Crop Yield

Fitting the Model and Checking Assumptions

The full ANOVA model involves the interaction terms μ_{jk}

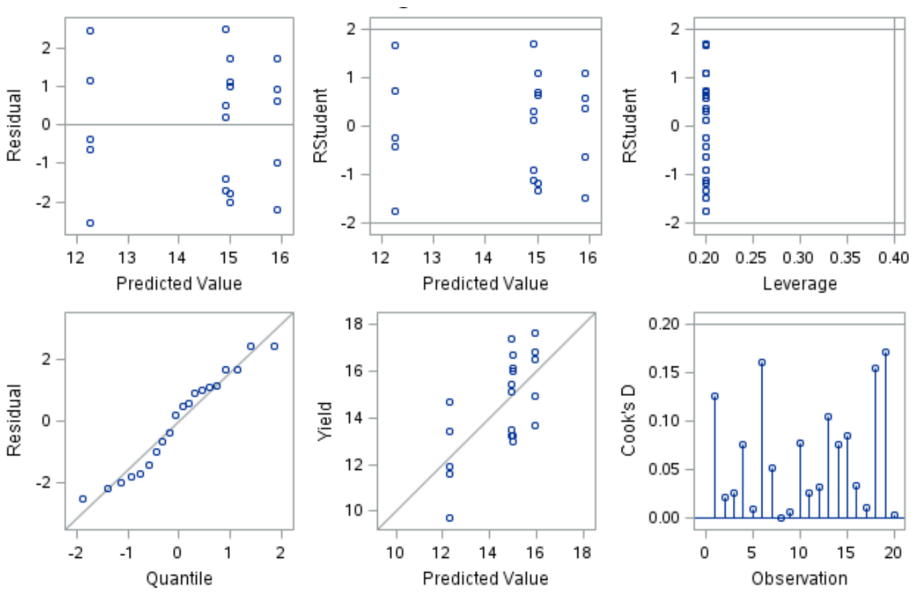
This encodes the potential for the mean response to vary nonlinearly

This complicates the interpretation (though not always disastrously, we will return to this at the end of these slides)

So, the first step is to test for a nonzero interaction

This, as always, requires a look at the assumptions

```
PROC GLM DATA = cropYield PLOTS = all;
    CLASS fert manure;
    MODEL yield = fert manure fert*manure;
RUN;
```



Testing for the Interaction

Suppose we have two factors A and B
(Say, A is type of fertilizer (J = 2) and B is type of manure (K = 2))

The interaction model would be ↔ $\mu\{Y|$ A, B$\}$ = A + B + A*B
ANOVA: $Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk}$
Regression: $Y_i = \beta_0 + \beta_1 Fertilizer + \beta_2 Plant + \beta_3 Fertilizer * Plant + \varepsilon_i$
(Note: in my opinion, the regression formulation makes the DF more transparent)

Source	DF	SS	MS	F	Pr > F
Fertilizer	J-1 = 1				
Species	K-1 = 1				
Fertilizer*Species	$(J-1)*(K-1)$ = 1				
Error	$n - J*K$ = 16				
Total	n - 1 = 19				

Testing for the Interaction

```
PROC GLM DATA = cropYield PLOTS = all;
    CLASS fert manure;
    MODEL yield = fert manure fert*manure;
RUN;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Fert	1	17.29800000	17.29800000	5.77	0.0288
Manure	1	15.84200000	15.84200000	5.28	0.0354
Fert*Manure	1	3.87200000	3.87200000	1.29	0.2727

Note: we will return to Type I vs III SS shortly

Here, no evidence for need of interaction term

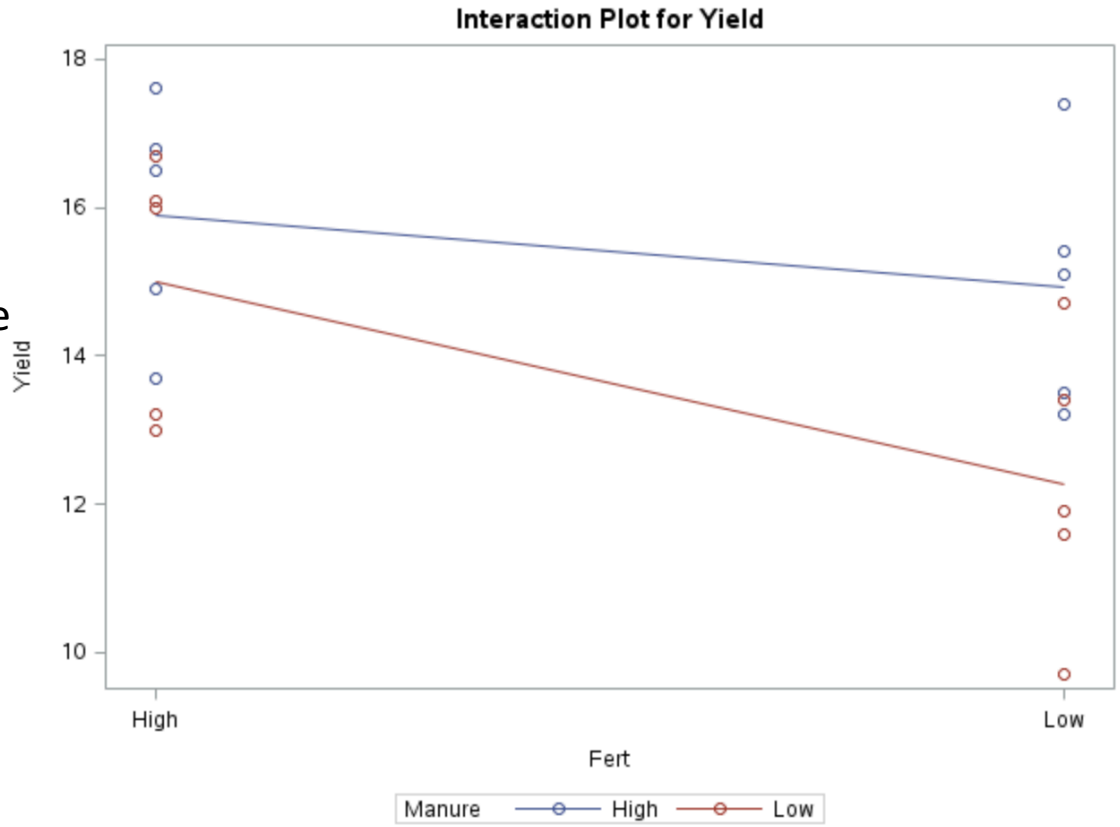
Source	DF	SS	MS	F	Pr > F
Fertilizer	$J-1 = 1$				
Species	$K-1 = 1$				
Fertilizer*Species	$(J-1)*(K-1) = 1$				
Error	$n - J*K = 16$				
Total	$n - 1 = 19$				

Profile Plot

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Fert	1	17.29800000	17.29800000	5.77	0.0288
Manure	1	15.84200000	15.84200000	5.28	0.0354
Fert*Manure	1	3.87200000	3.87200000	1.29	0.2727

Reminder: The test for interaction formalizes the question: Are the profile plot lines parallel?

Let's refit without that term



Fitting the Additive Model

```
PROC GLMMOD DATA = cropYield;
    CLASS fert manure;
    MODEL yield = fert manure;
RUN;
```

Parameter Definitions			
		CLASS Variable Values	
Column Number	Name of Associated Effect	Fert	Manure
1	Intercept		
2	Fert	High	
3	Fert	Low	
4	Manure		High
5	Manure		Low
6	Fert*Manure	High	High
7	Fert*Manure	High	Low
8	Fert*Manure	Low	High
9	Fert*Manure	Low	Low

Design Points										
		Column Number								
Observation Number	Yield	1	2	3	4	5	6	7	8	9
1	13.7	1	1	0	1	0	1	0	0	0
2	16.8	1	1	0	1	0	1	0	0	0
3	14.9	1	1	0	1	0	1	0	0	0
4	17.6	1	1	0	1	0	1	0	0	0
5	16.5	1	1	0	1	0	1	0	0	0
6	17.4	1	0	1	1	0	0	0	1	0
7	13.5	1	0	1	1	0	0	0	1	0
8	15.1	1	0	1	1	0	0	0	1	0
9	15.4	1	0	1	1	0	0	0	1	0
10	13.2	1	0	1	1	0	0	0	1	0
11	16.0	1	1	0	0	1	0	1	0	0
12	16.1	1	1	0	0	1	0	1	0	0
13	13.0	1	1	0	0	1	0	1	0	0
14	16.7	1	1	0	0	1	0	1	0	0
15	13.2	1	1	0	0	1	0	1	0	0
16	13.4	1	0	1	0	1	0	0	0	1
17	11.6	1	0	1	0	1	0	0	0	1
18	14.7	1	0	1	0	1	0	0	0	1
19	9.7	1	0	1	0	1	0	0	0	1
20	11.9	1	0	1	0	1	0	0	0	1

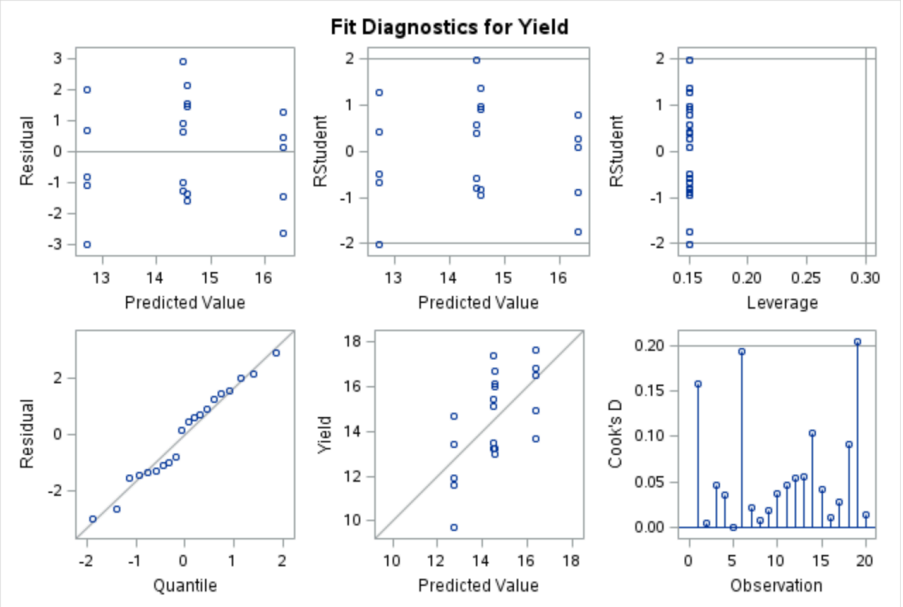
Fitting the Additive Model and Checking Assumptions

The additive model:

$$Y_{ijk} = \mu + \mu_j + \mu_k + \varepsilon_{ijk}$$

```
PROC GLM DATA = cropYield PLOTS = all;
    CLASS fert manure;
    MODEL yield = fert manure;
RUN;
```

This, as always, requires a look at the assumptions



Fitting the Additive Model

```
PROC GLM DATA = cropYield PLOTS = all;
    CLASS fert manure;
    MODEL yield = fert manure;
RUN;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Fert	1	17.29800000	17.29800000	5.67	0.0292
Manure	1	15.84200000	15.84200000	5.19	0.0359

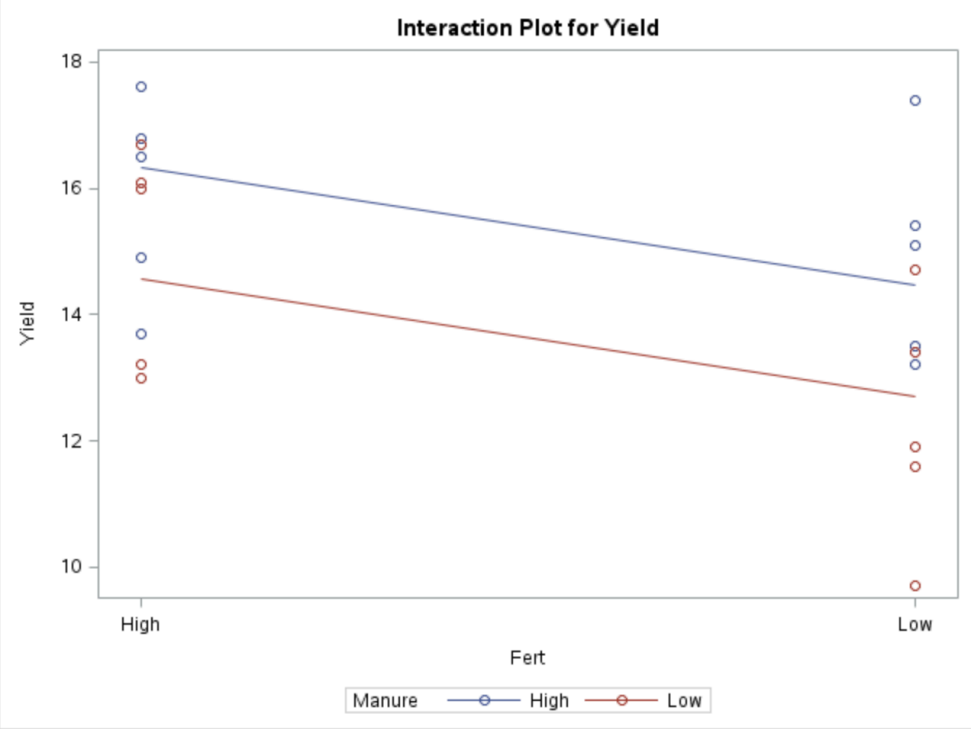
Here, there is evidence of a treatment effect for both Fertilizer and Manure

Profile Plot

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Fert	1	17.29800000	17.29800000	5.67	0.0292
Manure	1	15.84200000	15.84200000	5.19	0.0359

Important: The additive model fit will always have parallel lines, this provides no evidence against an interaction

What about estimating the effect?



Estimating the Additive Model

```
PROC GLM DATA = cropYield PLOTS = all;
    CLASS fert manure;
    MODEL yield = fert manure;
    LSMEANS fert manure / ADJUST = BON PDIFF CL;
RUN;
```

This produces least squares estimates or each level of each factor as well as a confidence interval for differences between the levels

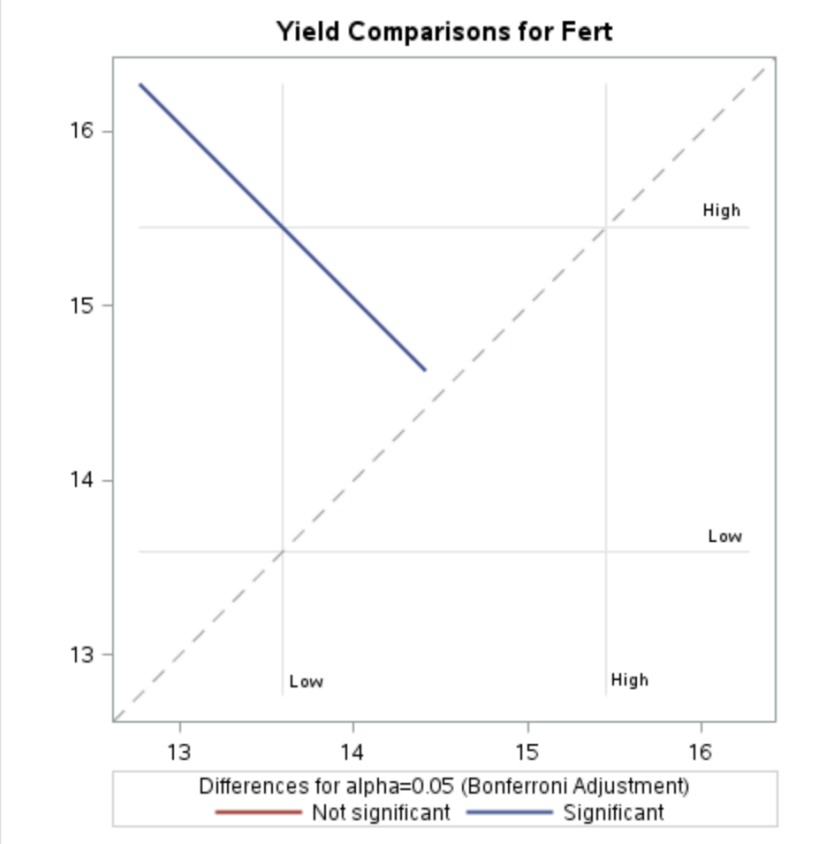
Fert	Yield LSMEAN	95% Confidence Limits	
High	15.450000	14.284569	16.615431
Low	13.590000	12.424569	14.755431

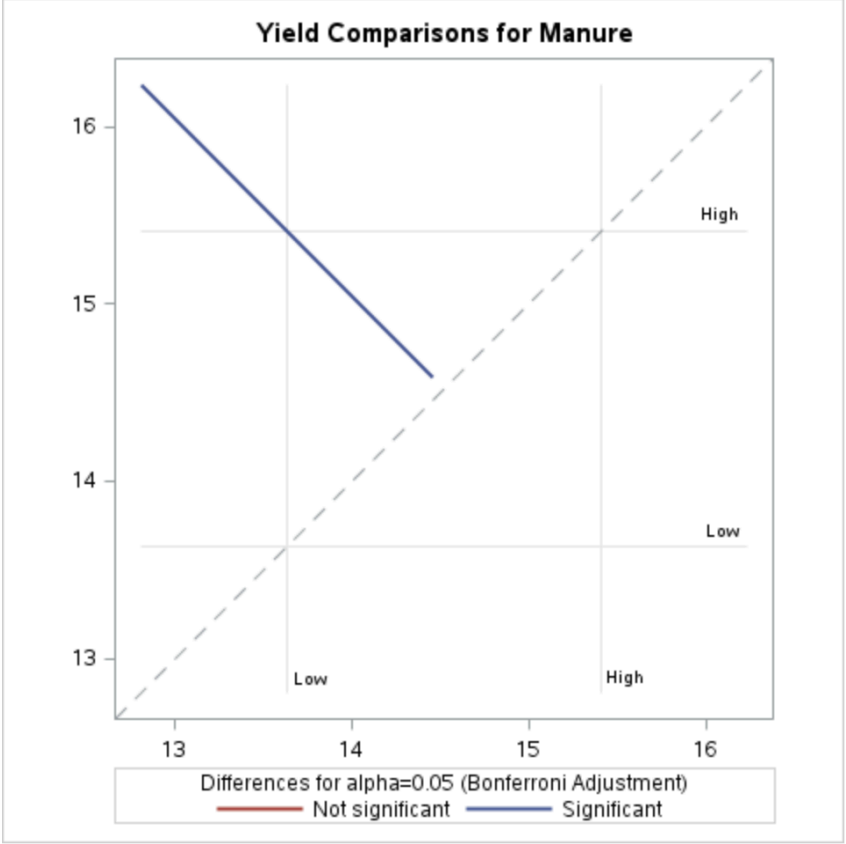
Least Squares Means for Effect Fert				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	1.860000	0.211832	3.508168

Manure	Yield LSMEAN	95% Confidence Limits	
High	15.410000	14.244569	16.575431
Low	13.630000	12.464569	14.795431

Least Squares Means for Effect Manure				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	1.780000	0.131832	3.428168

Estimating the Additive Model





What if There Was an Interaction?

(NOTE: I'M NOT INDICATING THIS ANALYSIS SHOULD BE DONE IN THIS CASE. I'M JUST SHOWING WHAT LSMEANS LOOKS LIKE WITH AN INTERACTION)

With an Interaction

```
PROC GLM DATA = cropYield PLOTS = all;
    CLASS fert manure;
    MODEL yield = fert manure fert*manure;
    LSMEANS fert manure fert*manure / ADJUST = BON PDIFF CL;
RUN;
```

This will estimate/compare the levels but take into account the interaction

With an Interaction

```
PROC GLM DATA = cropYield PLOTS = all;
    CLASS fert manure;
    MODEL yield = fert manure fert*manure;
    LSMEANS fert manure fert*manure / ADJUST = BON PDIFF CL;
RUN;
```

Fert	Manure	Yield LSMEAN	LSMEAN Number
High	High	15.9000000	1
High	Low	15.0000000	2
Low	High	14.9200000	3
Low	Low	12.2600000	4

Least Squares Means for effect Fert*Manure
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: Yield

i/j	1	2	3	4
1		1.0000	1.0000	0.0258
2	1.0000		1.0000	0.1417
3	1.0000	1.0000		0.1640
4	0.0258	0.1417	0.1640	

Fert	Manure	Yield LSMEAN	95% Confidence Limits	
High	High	15.900000	14.257928	17.542072
High	Low	15.000000	13.357928	16.642072
Low	High	14.920000	13.277928	16.562072
Low	Low	12.260000	10.617928	13.902072

With an Interaction